Tab A

CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope, with sufficient postage, addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, on

November 22, 2004
Date of Deposit

Tadashi Horie Reg. No. 40,437
Name of Applicant, Assignee or Registered Representative

Tadashi Horie
Signature

November 22, 2004
Date of Signature

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Appln. of: Toshihiko OBA
Appln. No.: 09/673,360
Filed: October 16, 2000
For: SPEECH TRANSFORMATION METHOD AND APPARATUS
Attorney Docket No: 11934/3

Examiner: NOLAN, Daniel A.
Art Unit: 2654

FOURTH SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT

In accordance with the duty of disclosure under 37 C.F.R. §1.56 and §§1.97-1.98, and more particularly in accordance with 37 C.F.R. §1.97(c), Applicant hereby cites the following reference:

DOCUMENT NUMBER Number-Kind Code (if known)	DATE	COUNTRY
H03-035296	02/15/1991	Japan

Applicant is enclosing Form PTO-1449 (one sheet), along with a copy of each listed reference for which a copy is required under 37 C.F.R. §1.98(a)(2). Reference E1 was cited in an Office Action issued in the counterpart Japanese application. A copy of the Office Action is also enclosed with this Statement. Reference E1 is in Japanese. Applicant is submitting a concise explanation of the relevant part of the reference. The

other references cited in the Office Action have already been disclosed in the previous Information Disclosure Statements.

Applicant respectfully requests the Examiner's consideration of the above reference and entry thereof into the record of this application.

By submitting this Statement, Applicant is attempting to fully comply with the duty of candor and good faith mandated by 37 C.F.R. §1.56. As such, this Statement is not intended to constitute an admission that any of the enclosed references, or other information referred to therein, constitutes "prior art" or is otherwise "material to patentability," as that phrase is defined in 37 C.F.R. §1.56(a).

For purposes of 37 C.F.R. §1.704(d), Applicant certifies that each item of information contained in this Statement was cited in a communication from a foreign patent office in a counterpart application, and that this communication was not received by any individual designated in 37 C.F.R. §1.56(c) more than thirty days prior to the filing of this Statement. As noted in the attached Office Action, the mailing date thereof is August 30, 2004.

Applicant hereby certifies under 37 C.F.R. §1.97(e)(1) that no item of information in this Statement was first cited in any communication from a foreign patent office in a counterpart foreign application more than three months prior to the filing of this Statement. Accordingly, Applicant has calculated no fee to be due in connection with the filing of this Statement. However, the Director is authorized to charge any fee deficiency associated with the filing of this Statement to a deposit account, as authorized in the Transmittal accompanying this Statement.

Respectfully submitted,

November 22, 2004
Date

Tadashi Horie
Tadashi Horie (Reg. No. 40,437)

BRINKS HOFER GILSON &LIONE

OIPE JC105 NOV 2 4 2004 PATENT & TRADEMARK OFFICE

RECEIVED
NOV 2 9 2004
Technology Center 2600

FORM PTO-1449	SERIAL NO. 09/673,360	CASE NO. 11934/3
LIST OF PATENTS AND PUBLICATIONS FOR APPLICANT'S FOURTH SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT	FILING DATE October 16, 2000	GROUP ART UNIT 2654
(use several sheets if necessary)	APPLICANT: Toshihiko OBA	

REFERENCE DESIGNATION **U.S. PATENT DOCUMENTS**

EXAMINER INITIAL		DOCUMENT NUMBER Number-Kind Code (if known)	DATE	NAME	CLASS/ SUBCLASS	FILING DATE
	E					
	E					
	E					
	E					
	E					
	E					

FOREIGN PATENT DOCUMENTS

EXAMINER INITIAL		DOCUMENT NUMBER Number-Kind Code (if known)	DATE	COUNTRY	CLASS/ SUBCLASS	Concise Explanation YES OR NO
	E1	H03-035296	02/15/1991	Japan		Yes
	E					
	E					
	E					
	E					

EXAMINER INITIAL		OTHER ART – NON PATENT LITERATURE DOCUMENTS (Include name of author, title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date page(s), volume-issue number(s), publisher, city and/or country where published.
	E	
	E	
	E	

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

特開平 3-35296

JP H03-35296

Patent Application No. H01-170230

Application H01 6/30

Title of the invention

A text speech synthesizer

An inventor

Kitou, Jungo

An applicant

SharpKK

[Example]

By the example which illustrated this devise, it is explained in detail.

FIG. 1 is a block diagram.

This is example of a text speech synthesizer of this devise.

The following is done in FIG. 1.

"A row of literal notation" is input into 31(Input unit according to literal notation)

"Example: The Japanese writing that Chinese character and kana are mixed" is input.

And it is sent out to 32(Analysis department according to literal notation).

"Morphological analysis, parsing and semantic analysis etc." of "a row of input literal notation"

32(Analysis department according to literal notation does this thing by consulting a dictionary.

It consults a dictionary same as so far, and, in this case, it is done.

"An analyzed morpheme" (word) and "a part of speech of each word" are output.

As for the case that "analyzed result is a word with activity", "grammer information of a conjugation" is output together.

32(Analysis department according to literal notation) comprises Fig4(21, 22, 23,24).

21) Analysis department of a morpheme, 22) Analysis department of a syntax ,23) Analysis department of semantics ,24)Dictionary

This is configured same as the conventional embodiment. This thing is omitted in FIG. 1.

"A control signal" (a designation of modification from relative difficulty word to an easy word) is input into 37 (control signal input part).

It is sent out to 38 (.The control section which changes relative difficulty word)

38 (The control section which changes relative difficulty word) analyzes an input

control signal.

And 38 (The control section which changes relative difficulty word) inputs various kinds of "relative difficulty word modification command" into 33 (The part which changes analysis of relative difficulty word).

"Command of modification of relative difficulty word" is following "control command etc".

Modification from relative difficulty word to an easy word is done (it is not done).

When "modification of relative difficulty word" is done ▼ "level of modification of relative difficulty word" is specified.

When "command of modification of relative difficulty word" from 38 (The control section which changes relative difficulty word) is input into 33(The part which changes analysis of relative difficulty word), the following is done.

Based on this command, "the relative difficulty word or homonym" of input command chapters is extracted.

Extracted word acts on on in "command of modification of relative difficulty word", and it is changed.

By means of the above-mentioned process, an input command section is changed in "easy sentence".

FIG. 4 is a detailed block diagram about 33 (The part which changes analysis of relative difficulty word).

This consists of 41 (The part which extracts relative difficulty word) and 42 (The part which changes relative difficulty word), and 43 (thesaurus) etc. 41 searches 43.

And "relative difficulty word, a homonym(an input command section)" are extracted.

42 (The part which changes relative difficulty word)searches 43 based on "command of modification of relative difficulty word" from 38(The control section which changes relative difficulty word) .

And (41 extract, : relative difficulty word, homonym) is changed in an easy word.

The following things are classed depending on a degree of difficulty to 43.

And these are stored.

They are "an easy single language of semantics the same as a homonym" and relative difficulty word.

On the occasion of modification of "relative difficulty word, a homonym", follows are done.

"Control command to specify level of relative difficulty word modification" output from 38 is based on, and the following is processed.

The following thing is chosen from "an easy single language".

It is a word of a degree of difficulty that there was in level of "modification of designated relative difficulty word".
Follows are done by the above being done. "Relative difficulty word, a homonym" are changed in an easy word.
It becomes "analysis result of a row of literal notation". This is sent out to 34 (The part which generates parameter of synthesized speech).
When a degree of difficulty is not changed, it seems to become follows.
When it is not changed a degree of difficulty of, it is written as follows.
"Command changing relative difficulty word" comprising "the control command that does not change relative difficulty word" is input from 38.
In such instances, 33 carries nothing out.
"Analysis result of a row of literal notation" in 32 is just sent out to 34 (The part which generates parameter of synthesized speech).
34 (The part which generates parameter of synthesized speech) that is "a section of a parameter of synthesized speech" is done as follows.
It is identified by 32 "analysis department of a row of literal notation" to control prosody.
It is processed as follows with "a modification command of relative difficulty word" from 38 (The control section which changes relative difficulty word).
Relative difficulty word and a homonym are changed in a simple word.
As thus described, by a changed "accent and syntactic structure of each word", the following process is done.
They are "clause when word did it in a chain reaction" and "an accent in expired paragraph" and "attachment of poise".
Even more particularly, time series of the following parameter as opposed to "synthesized speech corresponding to a vocalized sound voice" is got.
They are [duration, pitch pattern, power pattern, parameters of a special feature of a phoneme (coefficient of partial autocorrelation, line spectrum pair, Formant,etc.)].
35 (Speech synthesis part) are based on "parameter time series" for the speech synthesis, and follows are done.
Real "complex speech waveform" is generated, and it is output from 36 (synthesized speech output part).
When it is changed, in 33 (relative difficulty word analysis modification part), the following process is done by the word that "relative difficulty word and a homonym" are easy.
An operative example of "an input command section and output voice description" is

shown.

(example sentence 1)

Relative difficulty word is moved in easy sentence.

An input command section

The advance of {great strides} of science and engineering {made promote} development of industrial economy.

And social struture and institution were had an influence on {in various ways}.

Today's society is {revolutionized} by information and communication, and it {is referred to} an information society among other things.

Output voice description

The {fast} advancing of science and engineering {moved forward with} development of industrial economy.

And social structure and institution were had {all manner of} influences on.

Today's society {changes} by information and communication among other things ▼ {is said with} an information society.

(example sentence 2)

A homonym is explained in an easy word.

An input command section

An author {monopolizes} copyright.

When another person uses it, comprehension of an author must be got beforehand.

Output voice description

An author {monopolizes, owns alone} copyright.

When another person uses it, comprehension of an author must be got beforehand.

(example sentence 3)

A homonym is explained in an easy word.

An input command section

This proposal compiled {preliminary essay}.

This proposal compiled {preliminary essay, The thing which was described for trial}.

In example, a relative difficulty word and synonym extracted with "relative difficulty word extract department 41" are processed as follows to be seen in three example sentences.

It is changed in "an easy word" of a degree of difficulty of level of designated "relative difficulty word modification". By the process, it is output by synthesized speech. It is felt for a person to ask softly by these processes.

Therefore, it is easy to become hear very much. In addition, because there is not fear to be taken in wrong semantics ▼ description of information is transmitted enough by a

listener.
great strides : fast
make promote: move forward with
in various ways : all manner of
revolutionize : change
be reffered to : be said with
monopolize: own alone
preliminary essay : the thing which was described for trial
In that case of a homonym in "example sentence 2,3", it seems to become follows. Merely homonym is not moved in an easy word.
After having uttered in synthesized speech of a homonym ▼ synthesized speech of "another easy word which is the same semantics" is expressed in other words, and it is described. By these processes, a listener can know an original homonym. By it, a listener understands the nuance that an input command section is delicate.
In the exemplary embodiment, the following process is done based on a control signal from "control signal input part 37".
Various kinds of "relative difficulty word modification command" is input into "relative difficulty word modification control section 33" from "relative difficulty word modification control section 38".
"Relative difficulty word modification control section 33" do extract the same as "a language of relative difficulty or a homonym" from "a row of input literal notation".
From easy word configuration group of semantics the same as extracted "relative difficulty word or homonym", follows are chosen.
"A word of a degree of difficulty of level to be directed to" is chosen according to a control signal from "relative difficulty word modification control section 38". Relative difficulty word is changed in an easy word. An easy word is interposed in after "a homonym" alternatively.
1 : Relative difficulty word is changed in an easy word. 2 : After a homonym, an easy word corresponding to a homonym is interposed. It is handled for the case 1 or 2 as follows.
A voice composes it based on "a row of notation of an input command character", and it is output.Thus, it is processed as follows when "a word of the written language that it is difficult of semantics" difficult to hear is included in an input command section only in a voice.
It is output as "synthesized speech of spoken language to be composed of an easy word" to use by normal dialog.

Synthesized speech uttered by "a text speech synthesizer" reaches as follows. 1- 3). 1) It is felt to a person to ask softly, and it is easy to be heard very much. 2) It is taken in wrong semantics ▼ there is no possibility of it in it. 3) Description of information can be transmitted to a listener enough.

In the example, it is handled as follows. "A column of literal notation changed in an easy word" is based on by "relative difficulty word with a row of output literal notation", and it speechs

However, this devise is not limited to this. It may be done for the case a homonym similarly as follows. After relative difficulty word, "a column of the literal notation which interposed a selected easy word" is based on, and it utters to express it in other words in an easy word.

Brief description of drawings

Figure 1) A block diagram of example in this devise "text speech synthesizer"

Figure 2) A more detailed block diagram of "modification department in Fig.1 of relative difficulty word analysis"

Figure 3) A block diagram of a conventional "text speech synthesizer"

Figure 4) A more detailed block diagram of "analysis department of a row of literal notation"in Fig.1,3.

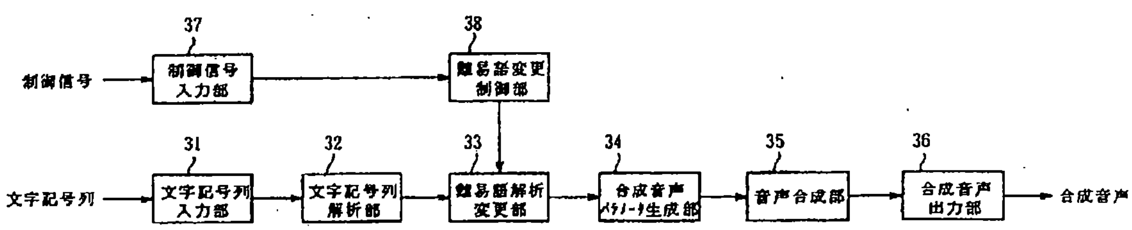
31) Input unit according to literal notation. 32) Analysis department according to literal notation. 33) The part which changes analysis of relative difficulty word. 34) The part which generates parameter of synthesized speech. 35) Speech synthesis part. 36) Synthesized speech output part 37) A control signal input part 38) The control section which changes relative difficulty word 41)The part which extracts relative difficulty word 42)The part which changes relative difficulty word 43) thesaurus

第3図は従来のテキスト音声合成装置のブロック図、第4図は第1図および第3図における文字記号列解析部のさらに詳細なブロック図である。

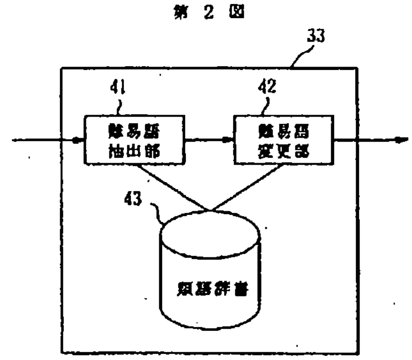
31…文字記号列入力部、
32…文字記号列解析部、
33…難易語解析変更部、
34…合成音声パラメータ生成部、
35…音声合成部、　36…合成音声出力部、
37…制御信号入力部、
38…難易語変更制御部、
41…難易語抽出部、　42…難易語変更部、
43…類語辞書。

特許出願人　シャープ株式会社
代理人　弁理士　青山　葆　ほか1名

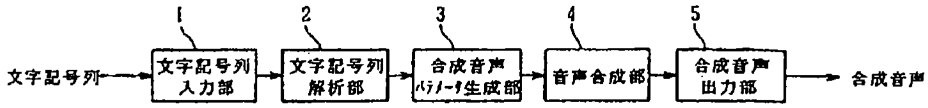
第1図



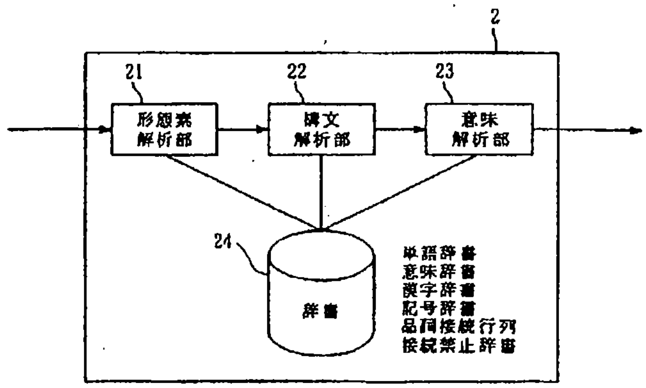
第2図



第 3 図


1
2
3
4
5
文字記号列
文字記号列
入力部
文字記号列
解析部
合成音声
パラメータ生成部
音声合成部
合成音声
出力部
合成音声

第 4 図


2
21
22
23
24
形態素
解析部
構文
解析部
意味
解析部
辞書
単語辞書
意味辞書
漢字辞書
記号辞書
品詞接続行列
接続禁止辞書

⑲日本国特許庁（JP）　⑪特許出願公開

⑫公開特許公報（A）　平3－35296

⑤Int. Cl.⁵　識別記号　庁内整理番号　㊸公開　平成3年(1991)2月15日
G 10 L 3/00　H　8622－5D

審査請求　未請求　請求項の数 2 （全7頁）

㊿発明の名称　テキスト音声合成装置

㉑特　願　平1－170230
㉒出　願　平1(1989)6月30日

⑫発明者　鬼頭　淳悟　大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内
⑫発明者　海木　延佳　大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内
⑪出願人　シャープ株式会社　大阪府大阪市阿倍野区長池町22番22号
⑭代理人　弁理士 青山 葆　外1名

明　細　書

1．発明の名称

テキスト音声合成装置

2．特許請求の範囲

（1）文章を構成する文字記号列の形態素，構文，意味等を文字記号列解析部で解析し、この解析結果に従って合成音声パラメータ生成部で音声合成用パラメータを生成し、この音声合成用パラメータに基づいて音声合成部で合成音声を生成して出力するテキスト音声合成装置において、

上記文字記号列解析部による解析結果に基づいて入力文章の中から難易語あるいは同音異義語を抽出し、この抽出した難易語あるいは同音異義語と同じ意味の平易な単語を選出し、選出された平易な単語を用いて入力文字記号列を変更し、この変更された文字記号列を上記合成音声パラメータ生成部に出力する難易語解析変更部を備えて、平易な単語から成る文体に変更された入力文章の合成音声を出力することを特徴とするテキスト音声合成装置。

（2）上記請求項1に記載のテキスト音声合成装置において、

入力文章の文字記号列を平易な文章の文字記号列に変更する際における変更のレベルを指定する指定信号を、上記難易語解析変更部に対して出力する難易語変更制御部を備える一方、

上記難易語解析変更部を、上記難易語変更制御部からの上記指定信号が入力された場合に、上記指定信号で指定された変更のレベルにある平易な単語を選出して入力文字記号列を変更するように成したことを特徴とするテキスト音声合成装置。

3．発明の詳細な説明

<産業上の利用分野>

この発明は、任意の文字記号列から成る入力文章を音声に変換するテキスト音声合成装置に関する。

<従来の技術>

従来、テキスト音声合成装置によって文字記号列から成る文章を合成音声に変換する際には、入力された任意の文字記号列に正しい読み，アクセ

ントおよびイントネーションを付加して音声合成用パラメータを生成する。そして、この生成された音声合成用パラメータに基づいて音声を合成して出力するようにしている。

すなわち、第3図において、文字記号列入力部1に文字記号列(例えば日本語漢字仮名混じり文)が入力されて文字記号列解析部2に送られる。そうすると、文字記号列解析部2は、後に詳述するようにして入力文字記号列の形態素(単語)解析，構文解析および意味解析等を行う。合成音声パラメータ生成部3は、韻律を制御するために、文字記号列解析部2における形態素解析によって同定された各単語のアクセントや構文構造から単語が連続した際の文節や呼気段落のアクセントやポーズの設定を行う。また、さらに発声音声に対応した合成単位に対する継続時間，ピッチパターン，パワーパターンおよび音韻特徴パラメータ(偏自己相関係数，線スペクトル対，ホルマント等)のパラメータ時系列を得る。そうすると、音声合成部4は上記音声合成用のパラメータ時系列に基づいて実際の合成音声波形を生成して合成音声出力部5から出力する。

第4図は文字記号列解析部2の更に詳細なブロック図である。文字記号列解析部2は形態素解析部21，構文解析部22，意味解析部23および辞書(単語辞書，意味辞書，漢字辞書，記号辞書，品詞接続行列，接続禁止辞書等)24によって構成されている。上記形態素解析部21は、文字記号列入力部1から入力された文字記号列を辞書(単語辞書，漢字辞書，記号辞書等)24を用いて形態素解析を行い単語を同定すると共に、同定した単語の品詞等の文法情報やアクセントを得る。上記構文解析部22は形態素解析部21によって同定された単語の構文を辞書(品詞接続行列および接続禁止辞書等)24を用いて決定する。また、上記意味解析部23は入力された文字記号列の意味を辞書(意味辞書等)24を用いて決定する。ただし、文字記号列解析部2は必ずしも形態素解析部21，構文解析部22，意味解析部23および辞書24によって構成する必要はない。すなわち、必要に応じて形態素解析部21，構文解析部22および辞書24、あるいは、形態素解析部21および辞書24のみによって構成してもよい。

すなわち、上記テキスト音声合成装置は、入力文章をそのまま合成音声によって読み上げるものである。

＜発明が解決しようとする課題＞

新聞や各種データベース等に見られる文章は目で読むことを前提にして作成されており、その殆どが書き言葉の文体によって表現されている。すなわち、通常耳にする話し言葉の文体とは大きく異なり、難易な単語や同音異義語が数多く含まれている。

しかしながら、上記テキスト音声合成装置においては、入力文章をそのまま合成音声によって読み上げるようになっているので、新聞や各種データベースを入力した場合には、書き言葉の文体での合成音声が出力される。したがって、その場合には通常聞き慣れている話し言葉の文体とはかなり異なり、出力される合成音声を聞く人には堅く感じられ、難易な単語や同音異義語が間違った意味に聞き取られる恐れがある。したがって、情報の出力手段としての機能を十分果たしているとはいえないという問題がある。

そこで、この発明の目的は、入力された文章を分かり易い自然な合成音声によって出力することができ、難易な単語や同音異義語が間違った意味に聞き取られる恐れのないテキスト音声合成装置を提供することにある。

＜課題を解決するための手段＞

上記目的を達成するため、この発明は、文章を構成する文字記号列の形態素，構文，意味等を文字記号列解析部で解析し、この解析結果に従って合成音声パラメータ生成部で音声合成用パラメータを生成し、この音声合成用パラメータに基づいて音声合成部で合成音声を生成して出力するテキスト音声合成装置において、上記文字記号列解析部による解析結果に基づいて入力文章の中から難易語あるいは同音異義語を抽出し、この抽出した難易語あるいは同音異義語と同じ意味の平易な単語

を選出し、選出された平易な単語を用いて入力文字記号列を変更し、この変更された文字記号列を上記合成音声パラメータ生成部に出力する難易語解析変更部を備えて、平易な単語から成る文体に変更された入力文章の合成音声を出力することを特徴としている。

また、上記テキスト音声合成装置は、入力文章の文字記号列を平易な文章の文字記号列に変更する際における変更のレベルを指定する指定信号を、上記難易語解析変更部に対して出力する難易語変更制御部を備える一方、上記難易語解析変更部を、上記難易語変更制御部からの上記指定信号が入力された場合に、上記指定信号で指定された変更のレベルにある平易な単語を選出して入力文字記号列を変更するように成すことが望ましい。

＜作用＞

この発明のテキスト音声合成装置において、文章を構成する文字記号列が文字記号列解析部に入力されて形態素解析，構文解析，意味解析等が実行される。そして、上記文字記号列解析部による解析結果が難易語解析変更部に入力される。そうすると、難易語解析変更部は、上記文字記号列解析部による解析結果に基づいて入力文章の中から難易語あるいは同音異義語を抽出し、この抽出された難易語あるいは同音異義語と同じ意味の平易な単語を選出する。そして、選出した平易な単語を用いて入力文字記号列が変更される。上記難易語解析変更部によって変更された文字記号列は合成音声パラメータ生成部に入力され、この変更後の文字記号列に従って音声合成用パラメータが生成される。そして、音声合成部によって、上記音声合成用パラメータに基づいて合成音声が生成されて出力される。したがって、平易な単語を用いて平易な文体に変更された入力文章の合成音声が出力される。

また、難易語変更制御部から、難易語変更のレベルを指定する指定信号が上記難易語解析変更部に対して出力されるようにする。そして、上記難易語変更制御部から上記指定信号が出力された場合に、この指定信号で指定されたレベルにある平易な単語が上記難易語解析変更部によって選出されて入力文字記号列が変更されるようにすれば、上記指定信号で指定されたレベルにある平易な単語を用いて変更された入力文章の合成音声を出力できる。

＜実施例＞

以下、この発明を図示の実施例により詳細に説明する。

第1図はこの発明のテキスト音声合成装置の一実施例を示すブロック図である。

第1図において、文字記号列入力部31に文字記号列(例えば日本語漢字仮名混じり文)が入力されて文字記号列解析部32に送出される。そうすると、文字記号列解析部32は、入力された文字記号列の形態素解析，構文解析および意味解析等を、上記従来例の場合と同様にして辞書を引いて行う。そして、解析された形態素(単語)および各単語の品詞を出力する。その際に、解析した結果が活用のある単語である場合には活用形等の文法情報も併せて出力する。ここで、第1図においては省略しているが、文字記号列解析部32は、上記従来例と同様に、第4図に示す形態素解析部21，構文解析部22，意味解析部23および辞書24によって構成されている。

一方、難易語から平易な単語への変更を指示するための制御信号が制御信号入力部37に入力されて難易語変更制御部38に送出される。そうすると、難易語変更制御部38は入力された制御信号を解析して種々の難易語変更指令を難易語解析変更部33に入力する。ここで、上記種々の難易語変更指令とは、難易語から平易な単語への変更を行う/行わないの制御指令および難易語変更を行う場合における難易語変更のレベルを指定する制御指令等である。

上記難易語解析変更部33は、難易語変更制御部38からの難易語変更指令が入力されると、この難易語変更指令に基づいて入力文章中の難易語あるいは同音異義語を抽出し、抽出した単語を上記難易語変更指令に従って変更することによって、入力文章を平易な文章に変更する。第4図は上記

難易語解析変更部33の更に詳細なブロック図であり、難易語抽出部41,難易語変更部42および類語辞書43等によって構成されている。上記難易語抽出部41は、類語辞書43を検索して、入力文章中に存在する難易語あるいは同音異義語を抽出する。上記難易語変更部42は、難易語変更制御部38からの難易語変更指令に基づいて類語辞書43を検索して、難易語抽出部41によって抽出された難易語や同音異義語を平易な単語に変更するのである。

ここで、上記類語辞書43には難易語あるいは同音異義語と同じ意味の平易な単語群が難易度に応じて分類されて格納されている。そして、難易語あるいは同音異義語の変更に際しては、難易語変更制御部38から出力される難易語変更のレベルを指定する制御指令に基づいて、指定された難易語変更のレベルに合った難易度の単語が平易な単語群の中から選択されるのである。

こうして、難易語や同音異義語が平易な単語に変更された文字記号列解析結果が合成音声パラメータ生成部34に送出される。一方、難易語変更を行わない場合(すなわち、難易語変更制御部38から難易語変更を行わない制御指令から成る難易語変更指令が入力された場合)には、難易語解析変更部33は何も実行しない。したがって、文字記号列解析部32における文字記号列解析結果がそのまま合成音声パラメータ生成部34に送出されるのである。

上記合成音声パラメータ生成部34は、韻律を制御するために、上記文字記号列解析部32によって同定され、さらに難易語変更制御部38からの難易語変更指令によって難易語や同音異義語が平易な単語に変更された各単語のアクセントや構文構造により、単語が連鎖した際の文節や呼気段落のアクセントやポーズの設定を行う。また、さらに発声音声に対応した合成単位に対する継続時間,ピッチパターン,パワーパターンおよび音韻特徴パラメータ(偏自己相関係数,線スペクトル対,ホルマント等)のパラメータ時系列を得る。そうすると、音声合成部35は上記音声合成用のパラメータ時系列に基づいて実際の合成音声波形を生成して合成音声出力部36から出力するのである。

次に、上記難易語解析変更部33において、難易語や同音異義語が平易な単語に変更された場合における、入力文章と出力音声内容との具体例を示す。

(例文1)難易語を平易な単語に置き換える。

入力文章

科学技術の長足の進歩は、産業経済の発達を促進すると共に、社会構造、制度にも種々の影響を及ぼした。とりわけ今日の社会は、情報及び、通信によって変革され、情報化社会と呼称される。

⇩

出力音声内容

科学技術の速い進歩は、産業経済の発達を押し進めると共に、社会構造、制度にもさまざまな影響を及ぼした。とりわけ今日の社会は、情報及び、通信によって変わり、情報化社会とよばれる。

(例文2)同音異義語を平易な単語で説明する。

入力文章

著作者は、著作権を専有しており、他人が利用する場合には事前に著作者の了解を得なければならない。

⇩

出力音声内容

著作者は、著作権を専有、独りで所有しており、他人が利用する場合には事前に著作者の了解を得なければならない。

(例文3)同音異義語を平易な単語で説明する。

入力文章

今回の提案は試論をまとめたものである。

⇩

出力音声内容

今回の提案は試論、試みに述べたものをまとめたものである。

上述の3つの例文に見られるように、本実施例においては、難易語抽出部41で抽出された難易語や同義語が

長足→ 速い

促進する→ 押し進める

種々の→ さまざまな
変革され → 変わり
呼称される → よばれる
専有 → 独りで所有
試論 → 試みに述べたもの

のように、指定された難易語変更のレベルに合った難易度における平易な単語に変更されて合成音声によって出力されるので、聞く人には柔らかく感じられて非常に聞き易いのである。また、間違った意味に取られる恐れがないので、十分に情報の内容が聞き手に伝達されるのである。

ここで、上記(例文2)および(例文3)における同音異義語の場合には、単に同音異義語を平易な単語に置き換えるのではなく、同音異義語の合成音声を発声した後に同じ意味の他の平易な単語の合成音声で言い換えて説明するようにしている。こうすることによって、聞き手は元の同音異義語を知ることができるので、入力文章の微妙なニュアンスをくみ取ることができるのである。

このように、本実施例においては、制御信号入力部37からの制御信号に基づいて難易語変更制御部38から種々の難易語変更指令が難易語変更制御部33に入力される。そうすると、難易語変更制御部33は、入力された文字記号列から難易語あるいは同音異義語を抽出し、上記抽出された難易語あるいは同音異義語と同じ意味の平易な単語群の中から、難易語変更制御部38からの制御指令に従って目的とするレベルの難易度の単語を選択し、上記難易語をこの平易な単語に変更するか、あるいは、同音異義語の後に平易な単語を挿入する。そして、この難易語が平易な単語に変更された、あるいは、同音異義語の後にその同音異義語に対応する平易な単語が挿入された入力文字記号列に基づいて、音声を合成して出力する。したがって、音声だけでは聞き取り取りにくいような意味の難しい書き言葉の単語が入力文章に含まれていても、通常の会話で用いられるような平易な単語から成る話し言葉の合成音声によって出力することができる。

すなわち、この発明のテキスト音声合成装置から発声される合成音声は、聞く人には柔らかく感じられて非常に聞き易く、間違った意味に取られる恐れがなく、情報の内容を十分に聞き手に伝達することができるのである。

上記実施例においては、入力文字記号列における難易語を平易な単語に変更した文字記号列に基づいて音声を合成するようにしている。しかしながら、この発明はこれに限定されるものではない。すなわち、同音異義語の場合と同様に、難易語の後に上記選択された平易な単語を挿入した文字記号列に基づいて平易な単語で言い換えるように発声させてもよい。

＜発明の効果＞

以上より明らかなように、この発明のテキスト音声合成装置は、難易語解析変更部によって、文字記号列解析部の解析結果に基づいて入力文章の中から難易語あるいは同音異義語を抽出し、この抽出した難易語あるいは同音異義語と同じ意味の平易な単語を選出し、この選出された平易な単語を用いて文字記号列を変更して合成音声パラメータ生成部に出力するようにしたので、書き言葉の文体の入力文章を分かり易い自然な話し言葉の文体による合成音声に変更して出力できる。したがって、難易語や同音異義語が間違った意味に聞き取られる恐れがない。

また、この発明のテキスト音声合成装置は、難易語変更制御部によって、入力文章の文字記号列を平易な文章の文字記号列に変更する際における変更のレベルを指定する指定信号を上記難易語解析変更部に出力し、この難易語解析変更部は、上記難易語変更制御部から上記指定信号が入力された場合に、上記指定信号で指定された変更レベルにある平易な単語を選出して入力文字記号列を変更するようにしたので、難易語変更の実施/不実施を指定でき、かつ、難易語変更のレベルを指定できる。

4、図面の簡単な説明

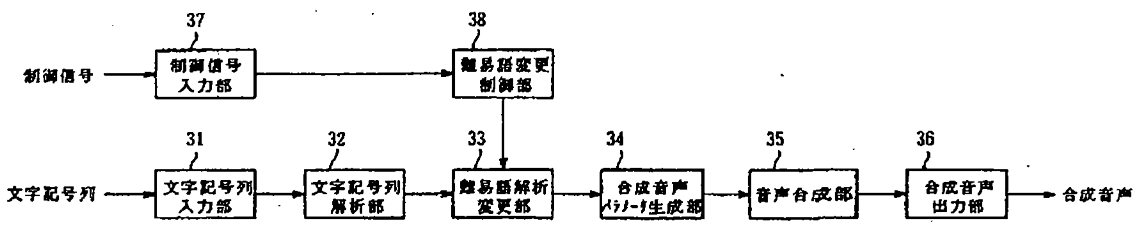
第1図はこの発明のテキスト音声合成装置における一実施例のブロック図、第2図は第1図における難易語解析変更部のさらに詳細なブロック図、

第3図は従来のテキスト音声合成装置のブロック図、第4図は第1図および第3図における文字記号列解析部のさらに詳細なブロック図である。

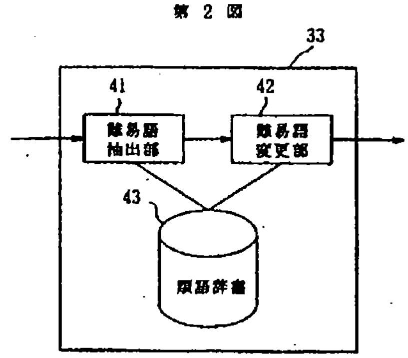
31…文字記号列入力部、
32…文字記号列解析部、
33…難易語解析変更部、
34…合成音声パラメータ生成部、
35…音声合成部、　36…合成音声出力部、
37…制御信号入力部、
38…難易語変更制御部、
41…難易語抽出部、　42…難易語変更部、
43…類語辞書。

特許出願人　シャープ株式会社
代理人　弁理士　青山　葆　ほか1名

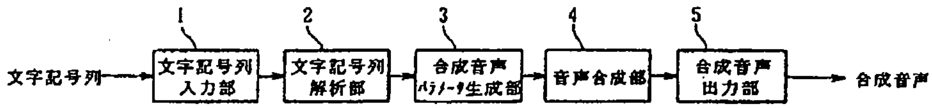
第1図



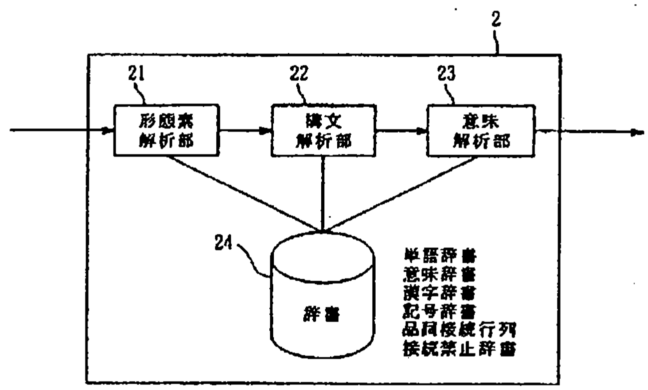
第2図



第 3 図


文字記号列
1
文字記号列入力部
2
文字記号列解析部
3
合成音声パラメータ生成部
4
音声合成部
5
合成音声出力部
合成音声

第 4 図


2
21
形態素解析部
22
構文解析部
23
意味解析部
24
辞書
単語辞書
意味辞書
漢字辞書
記号辞書
品詞接続行列
接続禁止辞書

Tab B

CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope, with sufficient postage, addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, on

June 15, 2005
Date of Deposit

Tadashi Horie Reg. No. 40,437
Name of Applicant, Assignee or Registered Representative

Signature

June 15, 2005
Date of Signature

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Appln. of: Toshihiko OBA

Appln. No.: 09/673,360

Filed: October 16, 2000

For: SPEECH TRANSFORMATION METHOD AND APPARATUS

Attorney Docket No: 11934/3

Examiner: NOLAN, Daniel A.

Art Unit: 2654

FIFTH SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT

In accordance with the duty of disclosure under 37 C.F.R. §1.56 and §§1.97-1.98, and more particularly in accordance with 37 C.F.R. §1.97(c), Applicant hereby cites the following reference:

FOREIGN DOCUMENT NUMBER Number-Kind Code (If known)	DATE	COUNTRY
H08-212228	08/20/1996	Japan

Applicant is enclosing Form PTO-1449 (one sheet), along with a copy of each listed reference for which a copy is required under 37 C.F.R. §1.98(a)(2). The attached reference is in Japanese. Applicant is submitting an English translation thereof. Applicant respectfully requests the Examiner's consideration of the above reference and entry thereof into the record of this application.

By submitting this Statement, Applicant is attempting to fully comply with the duty of candor and good faith mandated by 37 C.F.R. §1.56. As such, this Statement is not intended to constitute an admission that any of the enclosed references, or other information referred to therein, constitutes "prior art" or is otherwise "material to patentability," as that phrase is defined in 37 C.F.R. §1.56(a).

Applicant has calculated a processing fee in the amount of $180.00 to be due under 37 C.F.R. §1.17(p) in connection with the filing of this Statement. Applicant has enclosed a check covering this fee, or authorized charging the fee to a deposit account or credit card, as indicated in the Transmittal accompanying this Statement.

Respectfully submitted,

June 15, 2005
Date

Tadashi Horie (Reg. No. 40,437)

FORM PTO-1449	SERIAL NO. 09/673,360	CASE NO. 11934/3
LIST OF PATENTS AND PUBLICATIONS FOR APPLICANT'S FIFTH SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT	FILING DATE October 16, 2000	GROUP ART UNIT 2654
(use several sheets if necessary)	APPLICANT(S): Toshihiko OBA	

REFERENCE DESIGNATION **U.S. PATENT DOCUMENTS**

EXAMINER INITIAL		DOCUMENT NUMBER Number-Kind Code (if known)	DATE	NAME	CLASS/ SUBCLASS	FILING DATE
	F					
	F					
	F					
	F					
	F					
	F					

FOREIGN PATENT DOCUMENTS

EXAMINER INITIAL		DOCUMENT NUMBER Number-Kind Code (if known)	DATE	COUNTRY	CLASS/ SUBCLASS	TRANSLATION YES OR NO
	F1	H08-212228	08/20/1996	Japan		Yes
	F					
	F					
	F					

EXAMINER INITIAL		**OTHER ART – NON PATENT LITERATURE DOCUMENTS** (Include name of author, title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date page(s), volume-issue number(s), publisher, city and/or country where published.
	F	
	F	
	F	
	F	

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

H08-212228

Patent Laid-Open
[Publication Number]
Japanese Patent Laid-Open No. H08-212228 Laid-open date 1996/8/20
Title of the invention
"An abstract sentence making device" and "an abstract voice making device"
Application Number
Japanese Patent Application No. 7-16289
An applicant SANYO Electric
Inventor: Masayuki Iida, Kouji Tanaka, Masanori Miyatake

公開特許公報
【公開番号】
特開平８－２１２２２８　公開日１９９６/８/２０
発明の名称
「要約文作成装置」および「要約音声作成装置」
出願番号
特願平７－１６２８９
出願人　SANYO Electric
An example
An example of this invention is explained when taken with the drawing as follows.
[0038]
FIG. 1 is "an abstract sentence making device".
"An abstract sentence making device" is configured from "language analysis region 1" and "abstract sentence generator region 2".
A sentence is input into "language analysis department 1" as a character code line.
"Language analysis department 1" disassembles a sentence to "componentries such as a word, a phrase".
And 1 analyzes a sentence.
"Abstract sentence generator part 2" are based on an analysis result by "language analysis department 1", and it processes.
And 2 combines with an expensive componentry of importance, and an abstract sentence is generated.
[0039]
Figure 2-4 shows an operative example of "language analysis department 1"

respectively.

"Language analysis department 1" shown in FIG. 2 is configured from "morphological analysis department 11" comprising "morphological analysis dictionary 12]".

"Language analysis department 1" shown in FIG. 3 is configured from "morphological analysis region 11 comprising morphological analysis dictionary 12" and "syntax analyzer 13 comprising syntax rule 14".

"Language analysis department 1" shown in FIG. 4 is configured from "morphological analysis region 11 comprising morphological analysis dictionary 12" and "syntax analyzer 13 comprising syntax rule 14" and "semantic analyzer 15 comprising semantic dictionary 16".

[0040]

"Morphological analysis department 11" cut and bring down "a unit character string composing a sentence".

11 extracts "grammar information about each unit sentence character string".

Unit sentence character string is usually a word.

In addition, there is "a part of speech / an activity mode, etc.1" in grammar information.

As for the morphological analysis processing, an example of the algorithm is described as follows.

Therefore, this is well-known technology.

"Kouza Genzaino Gengo 7 (1984 Machine Process of Language: Makoto Nagao Sanseido ")

[0041]

"Morphological analysis dictionary 12" consist of memory means like ROM.

FIG. 5 shows an example for one part of an entry table installed in in "morphological analysis dictionary 12".

In this particular example, data to relate to "part of speech / information / practical use type" are stored by every word of "an entry".

Subject matter of "an entry" is expressed in a character code (a JIS code) line to show a word.

[0042]

About subject matter of "practical use type" of a word corresponding to "a verb, an adjectival noun", it seems to be follows.

It is stored by a practical use table installed in "morphological analysis dictionary 12".

In "a practical use model" of an entry table, "data to show the practical use table which should be referred" to in are stored.

An example of a practical use table is shown in FIG. 6.

[0043]
An example of an input sentence
While doing the device how our ancestors are various ▼ technology was accumulated.
(Japanese)
Watashitatachino,sosennha,samazamanakufuwo,sinagara, gijyutsuwo,chikusekisitekita.
In such case, it is processed as follows.
With "morphological analysis department 11", it seems to become the following from this sentence.
Watashitatachi, no,sosenn,ha,samazamana,kufu,wo,si,nagara, gijyutsu,wo,chikuseki,si,te,ki,ta.
A word is begun to talk about as above.
Information attached to a part of speech is analyzed about each word.
FIG. 7 shows an example of an analysis result of "morphological analysis department 11" as opposed to the example.
[0044]
When "language analysis department 1" is configured by means of "morphological analysis department 11" as shown in FIG. 2, it seems to become the following.
"Abstract sentence generator part 2" process based on an analysis result of "morphological analysis department 11" as follows.
A high word of importance is put together, and an abstract sentence is generated.
For example, a part of speech of a word is accepted, and, as for the importance of this case, it is decided as shown in next table 1.
[0045]
[table 1]
[0046]
"Abstract sentence generator part 2" hold an importance table shown in table 1.
Generation of an abstract sentence using a part of speech to all importance levels.
For example, this is set by manual operation.
When "high" was appointed as an importance level, it is processed as follows.
"A noun, a pronoun, a verb and the particle" which are a part of speech corresponding to importance level "high"
"Abstract sentence generator part 2" combine with these, and an abstract sentence is generated.
When "middle" was appointed as an importance level, it is processed as follows.
"A noun, a pronoun, a verb, a particle, an adjective and an adjectival noun" corresponding to "an importance level" (middle,high)

"Abstract sentence generator part 2" combine with these words, and an abstract sentence is generated.

[0047]

"Syntax analyzer 13" demand a syntactic structure of a statement based on "Morphological analysis department 11 analysis results" and "syntax rule 14".

In addition, syntax analysis processing is well known so that an example of algorithm is written as follows.

Example "Kouza Genzaino Gengo 7 (1984 Machine Process of Language: Makoto Nagao Sanseido "

[0048]

"Syntax rule 14" consist of memory means like ROM same as "morphological analysis dictionary 12".

To "syntax rule 14", an example is shown in FIG. 8.

"A combination state of a part of speech" and relation with a phrase

This is memorized in a table form.

Based on a part of speech of a word provided by means of "morphological analysis department 11", unification of a word is done.

[0049]

Based on a syntax rule as shown in FIG. 8, words are unified.

If "structure on the left hand side of FIG. 8" is discovered ▼ is defined as "a phrase of the right side".

"Noun phrase provided as a result of unification" is unified as "a noun on the left hand side of FIG. 8" more.

[0050]

Example sentense [We ....accumulated] "We" + "の" is unified in what is structure of a pronoun + particle, and, in 」, it is defined as a noun phrase of "our thing".

"Watashitachi" + "no" A pronoun + particle

It is unified, and it is defined as noun phrase of "Watashitachino".

"Sosen" + "ha" A noun + particle

It is unified, and it is defined as noun phrase of "Sosenha".

The noun phrase which was provided in this way

"Watashitachino" + [Sosenha A noun + noun

It is unified, and it is defined as noun phrase of "Watashitachino-sosenha".

[0051]

Example Sentense "Watashitachino.....Chikusekisitekita: Japanese"

A unification result as opposed to the above is shown in FIG. 9 along with an

integration process.
"Syntax analyzer 13" analyze it in an integration process as follows.
By "a particle, a conjunction, an attribute, a function to have such as a verb" included in a provided phrase by unification, 13 analyzes a syntactic structure.
[0052]
"Syntax analyzer 13" do the following decision process in an integration process more to be concrete when it is explained.
[0053]
(1)
"The nominative case, an objective case" of a sentence element are detected than "postpositional particle of function".
By means of existence of a verb, the predicate of a sentence element is detected.
"Watashtachino......Chikusekishitekita]"
By an entity of "postpositional particle of function:ha ", noun phrase "Sosenha" is judged to be subjective case.
By an entity of "postpositional particle of function:wo", noun phrase "Kufuuwo" and "Gijyutsu" are judged to be an objective case.
By an entity of a verb, phrasal verb(shi) and "Chikusekishitakita" are judged to be predicate.
[0054]
(2)
Based on estimate result from (1), it is handled as follows.
Main clause "Watsshitachinososenhakoudonagijyutsuwochikusekishitekita"
A subordination sentence
"Watashitachinososenhasamazamanakufuuwoshinagara"
It is determined that a sentence is composed of these.
In this particular example, it is distinguished from the main clause and a subordination sentence by connective particle "Nagara".
[0055]
When the shortest abstract sentence is made based on an analysis result of "syntax analyzer 13", it processes as follows.
From the nominative case of the main clause and an objective case and the predicate, it seems to become the following.
"Sosenha, gijyutsuwo tikusekisitekita"
[0056]
When "language analysis department 1" is configured by "morphological analysis

department 11" and "syntax analyzer 13" as shown in FIG. 3, it seems to become the following.
"Abstract sentence generator part 2" are based on "morphological analysis department 11" and an analysis result with "syntax analyzer 13", and it seems to become the following.
A high phrase of importance is put together, and an abstract sentence is generated.
[0057]
There are a lot of classifications for a phrase, but ▼ the nominative case, an objective case and the predicate do a frame of a sentence.
It is the most important phrase.
In addition, there are place status, clock time in other status.
For example, importance of a phrase is determined like next table 2.
[0058]
[table 2]
[0059]
"Abstract sentence generator part 2" hold "an importance table" as shown in table 2.
Generate abstract region by means of a phrase to all importance levels
For example, this is set by manual operation.
Among the main clause and subordination sentences, only the main clause or both is chosen.
This is set by manual operation.
[0060]
"Semantic analyzer 15" are handled based on "11 analysis results of 11 and 16" as follows.
Meanings such as "a word begun to talk about by means of 11" or "the phrase that words were unified" are analyzed.
Morphological analysis department : 11
Semantic dictionary : 16
[0061]
"Semantic dictionary 16" consist of memory means like ROM same as "morphological analysis dictionary 12".
To "semantic dictionary 16", "semantic information" is memorized every word of "an entry".
As for the thing, an example is shown in Fig.10
By means of "semantic dictionary 16", semantic information of a word is provided.
And, by this, semantic information of a passage including this word and a phrase can be

detected, too.

Semantic information of word "Kufuu" in example "Watashtachino.....chikusekishitekita]" is "Kagakugijyutsu".

It is hit subordination sentence "Watashitachinososenha samazamana kufuuwoshinagara" including this word percent semantic information "Kagakugijyutsu".

[0062]

When "language analysis department 1" is configured by "morphological analysis department 11" and "syntax analyzer 13" and "semantic analyzer 15" as shown in FIG. 4, it is processed as follows.

"Abstract sentence generator part 2" are based on "an analysis result with 11 and 13 and 15", and a high phrase of importance is put together, and an abstract sentence is generated.

"morphological analysis department ▼ 11" and "syntax analyzer ▼ 13" and "semantic analyzer ▼ 15 ":

[0063]

For example, it processes as follows when "gijyutsujhouhou" is specified as the semantic information that neutron importance is high by manual operation.

Neutron importance is determined principal clause a dependency statement "Watashitachinososenha samazamana kufuuwo shinagaramo"falling under this semantic information equally if high.

Thus, in this case, as for being similar, the each of nominative case and an objective case and the predicate of a subordination sentence are made to grapple with the main clause.

And an abstract sentence of "Sosenha kufuuwo shinagara gijyutsuwo tikusekisitekita: While an ancestor devises it ▼ a technique was accumulated" is made.

[0064]

In addition, there are politics economy, medicine, a law in "semantic information" other than technology.

These semantic information is appointed as the semantic information that neutron importance is high.

The abstract sentence that hit "a field of designated semantic information" with auto focusing is made by this.

[0065]

In each example, a word or an importance level of a phrase is set by manual operation.

"An importance level or importance of a subordination sentence" may be determined

depending on a set abstract rate automatically.
[0066]
With an abstract rate, "the ratio of length of a sentence of an abstract sentence" as opposed to "length of a sentence of the original" is said.
"Language analysis department 1" is composed of "morphological analysis department 11" and "syntax analyzer 13" as shown in FIG. 3.
For this case, it is processed as follows.
By "a combination with an importance level and importance of a subordination sentence", it is possible for a choice of four abstract levels like next table 3.
[0067]
[table 3]
[0068]
When a summary rate was set to 1/3, the process seems to be following.
Among "choices of four abstract levels", an abstract rate chooses the thing nearby to 1/3.
An abstract sentence is made by it.
[0069]
With the thing which "language analysis department 1" includes "morphological analysis department 11" and "syntax analyzer 13" and "semantic analyzer 15" in as shown in FIG. 4, it processes as follows.
Based on "designated important semantic information" , an entity of "a word, clause, a phrase" of high magnitude is added.
A shell abstract sentence is made among "choices of four abstract levels of table 3".
"As a result, the thing that an abstract rate becomes almost 1/3 most" is chosen.
An abstract sentence is made by these processing.
This choice is done automatically by working to summarize all choices.
[0070]
FIG. 11 shows an abstract voice making device.
This abstract voice making device comprises speech recognizer 21, abstract sentence making device 22, speech synthesis region 23.
[0071]
A voice is input into speech recognizer 21.
Speech recognizer 21 recognizes an input sound voice.
And 21 converts an input sound voice to a character code line.
Speech recognizer 21 does single syllable recognition.
Single tone clause is equivalent to kana 1 character.

Thus "a cf. single tone clause voice pattern memorized beforehand" and "a pattern of an input sound voice" are compared.
"Technique of stepless dynamic programming" is used for this comparison operation.
As a result of such a comparison operation, it is processed as follows.
There is "reference voice pattern" resembling a pattern of an input sound voice.
"Letter code column corresponding to this" is output along with similarity.
"Letter code column provided by means of speech recognizer 21" is sent to abstract sentence making device 22.
[0072]
Abstract sentence making device 22 processes as follows.
There is a componentry of a sentence expressed in "input letter code column".
22 determines importance of this.
22 combines with an expensive componentry of importance, and an abstract sentence is made.
FIG. 2, figure 3 or figure 4 (Summary statement implementation equipment 1) as 22 is used.
An abstract sentence is made with "abstract sentence making device 22".
This is sent to speech synthesis part 23 as a character code line.
[0073]
"Speech synthesis part 23" process as follows.
"Voice reference patern" is used, and "in need of summarized writing made with abstract sentence making device 22" is processed.
And it is converted into a voice, and it is output.
"Speech synthesis part 23" are composed of a speech synthesis by rule device.
A speech synthesis by rule device comprises the following.
It is a unit in kana 1 character
It "is a unit by a combination (around 2000) of a consonant / a vowel sound / a consonant"
"The phoneme memory which stored segment signal wave form of a voice" which assumed these a unit
"A segment signal waveform" corresponding to "input letter code column" is connected.
In the case of this connection processing, it is processed as follows.
"Representative accent and inflection (fundamental frequency variation) "are added to signal wave form based on "a convention corresponding to an array of a character code".
[0074]

FIG. 12 shows other abstract voice making devices.
This abstract voice making device comprises speech recognizer 21, abstract sentence making device 22, speech synthesis region 23 and quality of voice converter 24.
Speech recognizer 21, abstract sentence making device 22 and speech synthesis region 23 is the same as a thing shown in FIG. 11.
As thus described, the explanation is omitted.
[0075]
Quality of voice converter 24 processes as follows.
Based on "an interval of an input sound voice" (a voice input into an abstract voice making device) and "tone quality" (speech spectrum), it processes as follows.
"Quality of voice of a voice generated with speech synthesis part 23" is converted to "the sound quality that accepted quality of voice of an input sound voice".
Thus, it is it when "quality of voice of listing voice of a summary statement" is similar to "quality of voice of voice input by summary voice implementation equipment".
Thus, for example, an output voice becomes a feminine voice in the event of the voice that an input sound voice is feminine.
An input sound voice suffers from an output voice with a voice of an old man in the event of a voice of an old man.
"An output voice" depending on "sex, age" of an input sound voice is provided.
[0076]
FIG. 13 shows other abstract voice making devices more.
This abstract voice making device comprises buffer memory 31, speech recognizer 32, abstract sentence making device 33 and voice editing region 34.
"Speech recognizer 32 and abstract sentence making device 33" are the same as "speech recognizer 21 and abstract sentence making device 22" shown in FIG. 11.
[0077]
An input sound voice is accumulated to buffer memory 31.
And this is sent to speech recognizer 32 sequentially.
Speech recognizer 32 recognizes the input sound voice that has been sent from buffer memory 31.
And this converts a recognition result to a character code line, and it is output.
An address in buffer memory 31 of "a voice corresponding to a recognition result"
In doing so, these are put together, and it is output.
[0078]
Abstract sentence making device 33 determines "importance of a componentry of a sentence expressed in a character code line input from speech recognizer 32".

33 combines with an expensive componentry of importance, and an abstract sentence is made.
And, along with an address corresponding to "a componentry determined that neutron importance is high", a generated abstract sentence is sent to voice editing region 34.
[0079]
Voice editing department 34 is based on an input address, and the following is processed.
"A unit voice" corresponding to "each componentry composing an abstract sentence" is read from buffer memory 31.
This is connected to it.
Speech waveform depending on an abstract sentence is generated.
A "The important unit voice which is memorized to buffer memory 31 during an input sound voice"
B "Voice (summary voice) as opposed to a summary statement" "Voice as opposed to a summary statement" (summary voice)
A is connected, and B is made.
By the above-mentioned processing, quality of voice of a voice as opposed to an abstract sentence becomes approximately the same as quality of voice of an input sound voice.
[0080]
FIG. 14 shows other abstract voice making devices more.
This abstract voice making device comprises buffer memory 31, speech recognizer 32, abstract sentence making device 33, voice editing region 34 and prosody adjustment region 35.
Buffer memory 31, speech recognizer 32 and abstract sentence making device 33 is the same as a thing shown in FIG. 13.
Therefore, those explanation is omitted.
[0081]
Voice editing department 34 is based on an input address, and the following is processed.
"A unit voice" corresponding to "each componentry composing an abstract sentence" is read from buffer memory 31.
And it is connected.
And speech waveform depending on an abstract sentence is generated.
Generated speech waveform is sent to prosody regulation department 35.
A : Overall length of a continuation clock time of "each connected unit voice"

In doing so, A is sent to prosody regulation department 35 as incidental information, too.

[0082]

Prosody regulation department 35 processes as follows.

Joint of "each unit voice to compose a voice" edited with voice editing department 34

This is smoothed off by means of doing accent adjustment.

To prosody regulation department 35, the following is spent

Abstract speech waveform from voice editing department 34

Length of continuation time of each unit voice,

A character code line expressing an abstract sentence from abstract sentence generator device 33

[0083]

Fig.15 shows a configuration for prosody regulation department 35.

Accent section 41 processes by means of "accent dictionary 42 and accent transformational rule 43" as follows.

Accent information is extracted from the abstract sentence that has been sent from abstract sentence making device 33.

In a case of "Sosen", a part of "so" has an accent.

In addition, for extraction handling of this accent information, a morphological analysis and a syntax analysis are necessary.

Both analysis results with "morphological analysis department 11 and syntax analyzer 13" of abstract sentence making device 33 can be used.

Extracted accent information is sent to pace pattern section 44.

[0084]

"Pace pattern section 44" are based on "continuation length of time of each unit voice that has been sent from voice editing department 34", and the following is processed.

A pace pattern is generated so that a place with an accent becomes high.

There is technique represented for this generation method by "Fujisaki model".

[0085]

With pitch extraction department 45, a real pace pattern of speech waveform of "in need of summarized writing that has been sent from voice editing department 34" is extracted.

Various technique such as technique based on autocorrelation is known to a pitch extraction method.

[0086]

Interval converter 46 processes as follows.

"A pace pattern extracted with pitch extraction department 45" processes the following so that it is it with "a pace pattern generated by pace pattern section 44".
An interval of speech waveform of an abstract sentence is converted, and it is output.
Interval conversion technology is technology put to practical use in karaoke devices.
[0087]
An application of the "abstract sentence making device or abstract voice making device" is described.
[0088]
FIG. 16 shows the dictation system which an abstract sentence making device was applied to.
[0089]
"The audio signal which a voice was input into from microphone 101" or "an audio signal reproduced by tape reconstruction department 102" goes through A/D converter 103, and it is input in speech recognizer 112.
[0090]
For "an audio signal input into speech recognizer 112", voice input word processor processing is done by "speech recognizer 112 and document processing department 111 having a word processor function".
"A sentence comprising character code lines as opposed to the input sound voice that is this processing result" is stored by "a main memory (RAM) which is not illustrated, flow P disk 115, a memory means of the 116th class hard disk".
In addition, it is displayed to display 117 if necessary.
[0091]
As against a sentence stored by a memory means, abstract sentence making device 113 makes an abstract sentence automatically.
A made abstract sentence is stored by a main memory.
It is displayed with display 117 if necessary, and it is printed out with printer 106.
[0092]
In addition, if required, the following processing is done.
An abstract sentence made with abstract sentence making device 113 is converted to an audio signal by speech synthesis region 114.
It is seen off to loud speaker 105 through digital-to-analog converter 104, and voice output can be left afterwards.
[0093]
In addition, it is done as follows, and an abstract sentence can be made to fit into printing paper of the predetermined number of sheets.

In other words, for example, to document processing department 111, an order of the effect to make abstract is input into one piece of paper of A4 size.
Document processing department 111 is handled as follows.
It is processing in the abstract rate that can set with abstract sentence making device 113.
The abstract rate that an abstract sentence fits into in a range (number of characters) that one piece of paper of A4 size can describe
This is set to abstract sentence making device 113.
For such a parameter to set, there are paper size, a point size of a printer graphic, the paper number of sheets.
Number of characters of an abstract sentence is determined by appointing these parameters.
By such a function, one piece of minutes comprising abstract sentences can be made.
[0094]
In addition, OCR (belonging to a character recognition function) can be used as input means in a system of FIG. 16.
For this case, being similar make OCR recognize the meeting minutes of a blade a lot.
Based on this recognition result, one piece of minutes comprising abstract sentences can be made.
[0095]
FIG. 17 shows the example which applied an abstract voice making device to it at the time of high speed reproduction of VTR.
[0096]
Capstan servo circuit 201 is based on "a control signal from control signal head 202 and a velocity signal from capstan 203", and it processes as follows.
Capstan motor 204 is controlled so that travelling speed of videotape 205 becomes constant speed degree.
As "speed at the time of normal reproduction doubles travelling speed of videotape 205 at the time of two double speed reproduction", capstan motor 204 is controlled.
[0097]
Video head 206 reproduces a picture truck of videotape 205.
It is changed by predetermined order by head switching circuit 207, and video head 206 is output.
And it is converted into a picture signal with picture reflex circuit 208.
[0098]
Audio-head 209 reproduces an audio system truck of videotape 205.

A reproduced audio signal is sent to abstract voice making device 200.

[0099]

At high-speed reproduction, it is reproduced high speed a recorded "picture and voice" together by a videotape.

A high speed reproduced picture is displayed by a monitor.

A high speed reproduced voice is sent to abstract voice making device 200.

Abstract voice making device 200 generates "an abstract voice of utterance speed to be more late than utterance speed of a high-speed reproduction voice", and it is output.

For example, "an abstract voice of utterance speed of normal reproduction speed" is generated, and, at the time of two double speed reproduction, it is output.

[0100]

As thus described it is processed as follows when abstract voice making device 200 was applied to it at the time of high speed reproduction of VTR.

An abstract voice of slow utterance speed gets possible to be output than "utterance speed of a high-speed reproduction voice".

Therefore, an output voice at the time of high-speed reproduction is easy to become hear.

In addition, an abstract voice is output.

Therefore, it is processed as follows.

The number of words largely decreases than an original voice.

Therefore, it is hard to become generate that an abstract voice breaks off.

In addition, it is processed as follows by means of a disposition time to generate an abstract voice by an input sound voice.

When an output picture and a clock time gap between things of an output voice are outstanding, it is processed as follows.

By means of an image memory, it makes a video output delay.

And an output picture and the same period with an output voice are found.

[0101]

For an application of an abstract voice making device, there are a tape recorder, answering machine other than VTR.

Simplification of a voice and a saving of a tape can be planned.

[0102]

Application to answering machine of an abstract voice making device is explained.

Answering machine comprises a function to tape "a message of an opponent taken during going out".

Messages of the opponent who called during going out can be taped.

There is the following method to hear a taped message.
The first method
After return, answering machine is operated.
A taped message is revitalized, and it is heard.
The second method
Remote control assumes answering machine it from a going out former telephone, and a taped message is reproduced, and it is heard.
[0103]
For example, such an answering machine is connected to A/D converter 103 of a system of FIG. 16 and digital-to-analog converter 104.
And a reproduced audio signal is input into A/D converter 103 by means of "recording message playback equipment" in answering machine.
An abstract voice is made by an audio signal input into A/D converter 103 by a system of FIG. 16.
The audio signal is output from digital-to-analog converter 104.
[0104]
When a recording message is heard by the first method, it is processed as follows.
An abstract audio signal output from digital-to-analog converter 104 is output by answering machine.
When a recording message is heard by the second method, it is processed as follows.
An abstract audio signal output from digital-to-analog converter 104 is sent to going out ahead telephone through a phone line.
And it is output by going out ahead telephone.
In both methods, an abstract voice of a recording message is output by a telephone.
Message subject matter can be acquired in a short time.
In addition, in the event of the second method, cheapness of the phone line fee for use can be planned, too.
[0105]
[Effects of the Invention]
According to this invention, an abstract sentence can be made from a sentence automatically.
In addition, according to this invention, an abstract sentence can be made from an input voice automatically.
In addition, according to this invention, an abstract sentence is made from a sentence automatically.
A voice corresponding to a made abstract sentence can be output.

In addition, according to this invention, an abstract sentence is made from an input voice automatically.
A voice corresponding to a made abstract sentence can be output.
In addition, according to this invention, processing equal to or less than it is done at high-speed reproduction of playback equipment such as VTR.
An abstract voice of normal speed is made from a high speed reproduced voice, and it can be output.

Brief description of drawings 】
[FIG. 1]
It is a block diagram to show framing of an abstract sentence making device in.
[FIG. 2]
It is a block diagram showing constitution of language analysis department.
[FIG. 3]
It is a block diagram showing an example other than language analysis department.
[FIG. 4]
It is a block diagram showing other examples more of language analysis department.
[FIG. 5]
It is a schematic block diagram to show an example of an entry table in morphological analysis dictionary.
[FIG. 6]
It is a schematic block diagram to show an example of a practical use table in morphological analysis dictionary.
[FIG. 7]
It is a schematic block diagram to show a morphological analysis result.
[FIG. 8]
It is a schematic block diagram to show an example of a syntax rule.
[FIG. 9]
It is a schematic block diagram to show a syntax analysis result and an example of the process.
[Fig.10]
It is a schematic block diagram to show an example of subject matter of semantic dictionary.
[FIG. 11]
It is a block diagram to show framing of an abstract voice making device in.
[FIG. 12]

It is a block diagram showing an example other than an abstract voice making device.
[FIG. 13]
It is a block diagram to show other examples in more of an abstract voice making device.
[FIG. 14]
It is a block diagram to show other examples in more of an abstract voice making device.
[FIG. 15]
It is a block diagram showing constitution of prosody regulation department.
[FIG. 16]
It is a block diagram showing the dictation system which applied an abstract of the invention making device.
[FIG. 17]
It is a block diagram showing the application which applied an abstract of the invention making device to VTR.
[Denotation of Reference Numerals]
One language analysis department
Two abstract sentence generator part
11 morphological analyses department
12 morphological analysis dictionary
13 syntax analyzers
14 syntax rules
15 semantic analyzers
16 semantic dictionary
21, 32
A speech recognizer
22, 33, 113
An abstract sentence making device
23 speech synthesis part
24 quality of voice converters
31 buffer memories
34 voice editing department
35 prosody regulation department
200 abstract voice making devices
Table 1 [0045]
A part of speech

A noun
A pronoun
A verb
A particle
An adjective
An adjectival noun
An adverb
Importance
It is high
It is high
It is high
It is high
The inside
The inside
It is low

Table 2 [0058]
A phrase
The nominative case
An objective case
The predicate
Other status
Importance
It is high
It is high
It is high
It is low

Table 3 [0067]
A main clause / subordination sentence
Only one main clause
Only two main clauses
Three subordination sentences are included
Four subordination sentences are included
The pitch of a phrase (status)
1 is high

2 is low
3 is high
4 is low

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（Ａ）　(11)特許出願公開番号

特開平8－212228

(43)公開日　平成8年(1996) 8月20日

(51)Int.Cl.⁶	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/28					
Ｇ１０Ｌ　3/00	５６１　Ｇ				
		8420－5L	Ｇ０６Ｆ	15/ 38	R
		8420－5L			V

審査請求　未請求　請求項の数21　ＯＬ　（全 14 頁）

(21)出願番号　特願平7－16289

(22)出願日　平成7年(1995) 2月2日

(71)出願人　000001889
三洋電機株式会社
大阪府守口市京阪本通2丁目5番5号

(72)発明者　飯田　正幸
大阪府守口市京阪本通2丁目5番5号　三洋電機株式会社内

(72)発明者　田中　浩司
大阪府守口市京阪本通2丁目5番5号　三洋電機株式会社内

(72)発明者　宮武　正典
大阪府守口市京阪本通2丁目5番5号　三洋電機株式会社内

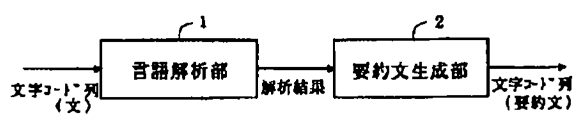
(74)代理人　弁理士　香山　秀幸

(54)【発明の名称】　要約文作成装置および要約音声作成装置

(57)【要約】

【目的】　この発明は、文から自動的に要約文を作成することができる要約文作成装置を提供することを目的とする。

【構成】　要約文作成装置において、文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成する。



1

【特許請求の範囲】

【請求項１】　文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成する要約文作成装置。

【請求項２】　文を構成要素に分解して文を解析する言語解析手段、および言語解析手段による解析結果に基づいて、重要度の高い構成要素を組み合わせて要約文を生成する要約文生成手段、

を備えている要約文作成装置。

【請求項３】　言語解析手段が、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する形態素解析手段からなり、

要約文生成手段が、形態素解析手段の解析結果に基づいて、重要度の高い単位文字列を組み合わせて要約文を生成する請求項２に記載の要約文作成装置。

【請求項４】　言語解析手段が、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する形態素解析手段および形態素解析手段の解析結果に基づいて、文法に則り文の構文構造を求める構文解析手段からなり、

要約文生成手段が、形態素解析手段および構文解析手段の解析結果に基づいて、単位文字列および／または単位文字列が統合されてなる単位文字列統合体のうち、重要度の高いものを組み合わせて要約文を生成する請求項２に記載の要約文作成装置。

【請求項５】　言語解析手段が、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する形態素解析手段、形態素解析手段の解析結果に基づいて、文法に則り文の構文構造を求める構文解析手段および形態素解析手段によって切り出された単位文字列および／または単位文字列が統合されてなる単位文字列統合体の意味を解析する意味解析手段からなり、

要約文生成手段が、形態素解析手段、構文解析手段および意味解析手段の解析結果に基づいて、単位文字列および／または単位文字列統合体のうち、重要度の高いものを組み合わせて要約文を生成する請求項２に記載の要約文作成装置。

【請求項６】　入力音声を認識して、入力音声を文字コード列に変換する音声認識手段、および文字コード列で表される文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成する要約文生成手段、

を備えている要約文作成装置。

【請求項７】　文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成する要約文生成手段、および生成された要約文の文字コード列に応じた音声を生成する音声合成手段、

を備えている要約音声作成装置。

【請求項８】　入力音声を認識して、入力音声を文字コード列に変換する音声認識手段、

2

文字コード列で表される文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成する要約文生成手段、および生成された要約文の文字コード列に応じた音声を生成する音声合成手段、

を備えている要約音声作成装置。

【請求項９】　入力音声を認識して、入力音声を文字コード列に変換する音声認識手段、

文字コード列で表される文を構成要素に分解して文を解析する言語解析手段、

言語解析手段による解析結果に基づいて、重要度の高い構成要素を組み合わせて要約文を生成する要約文生成手段、および生成された要約文の文字コード列に応じた音声を生成する音声合成手段、

を備えている要約音声作成装置。

【請求項１０】　言語解析手段が、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する形態素解析手段からなり、

要約文生成手段が、形態素解析手段の解析結果に基づいて、重要度の高い単位文字列を組み合わせて要約文を生成する請求項９に記載の要約音声作成装置。

【請求項１１】　言語解析手段が、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する形態素解析手段および形態素解析手段の解析結果に基づいて、文法に則り文の構文構造を求める構文解析手段からなり、

要約文生成手段が、形態素解析手段および構文解析手段の解析結果に基づいて、単位文字列および／または単位文字列が統合されてなる単位文字列統合体のうち、重要度の高いものを組み合わせて要約文を生成する請求項９に記載の要約音声作成装置。

【請求項１２】　言語解析手段が、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する形態素解析手段、形態素解析手段の解析結果に基づいて、文法に則り文の構造を求める構文解析手段および形態素解析手段によって切り出された単位文字列および／または単位文字列が統合されてなる単位文字列統合体の意味を解析する意味解析手段からなり、

要約文生成手段が、形態素解析手段、構文解析手段および意味解析手段の解析結果に基づいて、単位文字列および／または単位文字列統合体のうち、重要度の高いものを組み合わせて要約文を生成する請求項９に記載の要約音声作成装置。

【請求項１３】　音声合成部で生成された音声の声質を、入力音声の声質に応じた音質に変換する手段を備えている請求項７、８、９、１０、１１および１２のいずれかに記載の要約音声作成装置。

【請求項１４】　入力音声を記憶する記憶手段、

入力音声を認識して、入力音声を文字コード列に変換する音声認識手段、

文字コード列で表される文の構成要素の重要度を判定

し、重要度の高い構成要素を組み合わせて要約文を生成する要約文生成手段、ならびに生成された要約文を構成する各構成要素に対応する単位音声を記憶手段から読み出して、要約文に応じた音声を編集する音声編集手段、を備えている要約音声作成装置。

【請求項１５】　入力音声を記憶する記憶手段、
入力音声を認識して、入力音声を文字コード列に変換する音声認識手段、
文字コード列で表される文を構成要素に分解して文を解析する言語解析手段、
言語解析手段による解析結果に基づいて、重要度の高い構成要素を組み合わせて要約文を生成する要約文生成手段、ならびに生成された要約文を構成する各構成要素に対応する単位音声を記憶手段から読み出して、要約文に応じた音声を編集する音声編集手段、
を備えている要約音声作成装置。

【請求項１６】　言語解析手段が、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する形態素解析手段からなり、
要約文生成手段が、形態素解析手段の解析結果に基づいて、重要度の高い単位文字列を組み合わせて要約文を生成する請求項１５に記載の要約音声作成装置。

【請求項１７】　言語解析手段が、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する形態素解析手段および形態素解析手段の解析結果に基づいて、文法に則り文の構造を求める構文解析手段からなり、
要約文生成手段が、形態素解析手段および構文解析手段の解析結果に基づいて、単位文字列および／または単位文字列が統合されてなる単位文字列統合体のうち、重要度の高いものを組み合わせて要約文を生成する請求項１５に記載の要約音声作成装置。

【請求項１８】　言語解析手段が、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する形態素解析手段、形態素解析手段の解析結果に基づいて、文法に則り文の構造を求める構文解析手段および形態素解析手段によって切り出された単位文字列および／または単位文字列が統合されてなる単位文字列統合体の意味を解析する意味解析手段からなり、
要約文生成手段が、形態素解析手段、構文解析手段および意味解析手段の解析結果に基づいて、単位文字列および／または単位文字列統合体のうち、重要度の高いものを組み合わせて要約文を生成する請求項１５に記載の要約音声作成装置。

【請求項１９】　音声編集手段で編集された音声を構成する各単位音声の繋ぎ目をなめらかにする韻律調整手段を備えている請求項１４、１５、１６、１７および１８のいずれかに記載の要約音声作成装置。

【請求項２０】　少なくとも音声が記録された記録媒体を、標準再生速度より高速で再生する手段、

高速再生された音声を認識して、入力音声を文字コード列に変換する音声認識手段、
文字コード列で表される文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成する手段、ならびに生成された要約文の文字コード列に応じた、高速再生音声の発声速度より遅い発声速度の音声を生成して出力する音声合成手段、
を備えている要約音声作成装置。

【請求項２１】　映像と音声とが対応づけられて記録された記録媒体を、標準再生速度より高速で再生する手段、
高速再生された音声を認識して、入力音声を文字コード列に変換する音声認識手段、
文字コード列で表される文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成する手段、
生成された要約文の文字コード列に応じた、高速再生音声の発声速度より遅い発声速度の音声を生成する音声合成手段、ならびに高速再生された映像と音声合成手段によって生成された要約文に対する音声とを出力する出力手段、
を備えている映像・音声処理装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】この発明は、要約文作成装置、要約音声作成装置および映像・音声処理装置に関する。

【０００２】

【従来の技術】たとえば、ビデオテープレコーダ（ＶＴＲ）において、２倍速再生等の高速再生を行うと出力音声速度も標準音声速度の２倍となり、出力音声が聴き取りにくくなる。そこで、２倍速再生された音声を格納する音声メモリを設け、音声メモリの書き込み／読み出し速度を制御することにより、２倍速再生時に、音声を標準速度で出力させて、出力音声を聴き取り易くする技術がすでに開発されている。

【０００３】

【発明が解決しようとする課題】音声メモリの書き込み／読み出し速度を制御して、２倍速再生時に、音声を標準速度で出力させる方法においては、入力音声の半分が削除されてしまう。高速再生時に、内容理解に有用な音声が削除される割合を少なくするために、本出願人は、次のような話速変換装置を発明した。

【０００４】つまり、高速再生された音声信号のうち、無音区間を検出して無音区間を削除する。そして、高速再生された音声信号のうち、音声区間の信号に対して時間軸圧縮伸長処理を行って音声メモリに記憶させる。この場合、高速再生音声の発生速度よりも出力音声速度が遅くなるように圧縮率が設定される。そして、音声メモリに記憶された音声データを順次出力していく。この方法においても、音声メモリに書き込まれているが読み出

されていないデータ量が音声メモリの容量を越えると、音声メモリに蓄積されている音声データは削除され、削除された音声データは出力されなくなる。

【0005】本出願人は、ＶＴＲの高速再生時において、内容を把握できかつ音声の途切れのない出力音声を得るために、高速再生された音声からその要約文に相当する要約音声を生成し、高速再生音声の発生速度より遅い発生速度で出力することを着想した。この発明は、上記着想に基づいてなされたものである。

【0006】この発明は、文から自動的に要約文を作成することのできる要約文作成装置を提供することを目的とする。

【0007】この発明は、入力された音声から、自動的に要約文を作成することのできる要約文作成装置を提供することを目的とする。

【0008】この発明は、文から自動的に要約文を作成し、作成した要約文に対応する音声を出力できる要約音声作成装置を提供することを目的とする。

【0009】この発明は、入力された音声から、自動的に要約文を作成し、作成した要約文に対応する音声を出力できる要約音声作成装置を提供することを目的とする。

【0010】この発明は、ＶＴＲ等の再生装置の高速再生時において、高速再生された音声から標準速度の要約音声を作成して出力することができる要約音声作成装置を提供することを目的とする。

【0011】

【課題を解決するための手段】この発明による第１の要約文作成装置は、文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成することを特徴とする。ここで文の構成要素とは、文を構成する単語、節、句等をいう。

【0012】この発明による第２の要約文作成装置は、文を構成要素に分解して文を解析する言語解析手段および言語解析手段による解析結果に基づいて、重要度の高い構成要素を組み合わせて要約文を生成することを特徴とする。

【0013】この発明による第３の要約文作成装置は、入力音声を認識して、入力音声を文字コード列に変換する音声認識手段、および文字コード列で表される文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成することを特徴とする。

【0014】この発明による第１の要約音声作成装置は、文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成する要約文生成手段、および生成された要約文の文字コード列に応じた音声を生成する音声合成手段を備えていることを特徴とする。

【0015】この発明による第２の要約音声作成装置は、入力音声を認識して、入力音声を文字コード列に変換する音声認識手段、文字コード列で表される文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成する要約文生成手段、および生成された要約文の文字コード列に応じた音声を生成する音声合成手段を備えていることを特徴とする。

【0016】この発明による第３の要約音声作成装置は、入力音声を認識して、入力音声を文字コード列に変換する音声認識手段、文字コード列で表される文を構成要素に分解して文を解析する言語解析手段、言語解析手段による解析結果に基づいて、重要度の高い構成要素を組み合わせて要約文を生成する要約文生成手段、および生成された要約文の文字コード列に応じた音声を生成する音声合成手段を備えていることを特徴とする。

【0017】上記第１～第３の要約音声作成装置において、音声合成部で生成された音声の声質を、入力音声の声質に応じた音質に変換する手段を設けることが好ましい。

【0018】この発明による第４の要約音声作成装置は、入力音声を記憶する記憶手段、入力音声を認識して、入力音声を文字コード列に変換する音声認識手段、文字コード列で表される文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成する要約文生成手段、ならびに生成された要約文を構成する各構成要素に対応する単位音声を記憶手段から読み出して、要約文に応じた音声を編集する音声編集手段を備えていることを特徴とする。

【0019】この発明による第５の要約音声作成装置は、入力音声を記憶する記憶手段、入力音声を認識して、入力音声を文字コード列に変換する音声認識手段、文字コード列で表される文を構成要素に分解して文を解析する言語解析手段、言語解析手段による解析結果に基づいて、重要度の高い構成要素を組み合わせて要約文を生成する要約文生成手段、ならびに生成された要約文を構成する各構成要素に対応する単位音声を記憶手段から読み出して、要約文に応じた音声を編集する音声編集手段を備えていることを特徴とする。

【0020】上記第４または第５の要約音声作成装置において、音声編集手段で編集された音声を構成する各単位音声の繋ぎ目をなめらかにする韻律調整手段を設けることが好ましい。

【0021】上記第２の要約文作成装置、上記第３の要約音声作成装置または上記第５の要約音声作成装置における言語解析手段および要約文生成手段としては次のようなものが用いられる。

【0022】（１）　言語解析手段としては、たとえば、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する形態素解析手段からなるものが用いられる。この場合には、要約文生成手段としては、形態素解析手段の解析結果に基づいて、重要度の高い単位文字列を組み合わせて要約文を生成するものが用いられる。ここで、単位文字列は、たとえば、単語をい

う。

【0023】（2）　言語解析手段としては、たとえば、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する形態素解析手段および形態素解析手段の解析結果に基づいて、文法に則り文の構文構造を求める構文解析手段からなるものが用いられる。この場合には、要約文生成手段としては、形態素解析手段および構文解析手段の解析結果に基づいて、単位文字列および／または単位文字列が統合されてなる単位文字列統合体のうち、重要度の高いものを組み合わせて要約文を生成するものが用いられる。ここで、単位文字列統合体とは、たとえば、単語が統合された節、句等をいう。

【0024】（3）　言語解析手段としては、たとえば、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する形態素解析手段、形態素解析手段の解析結果に基づいて、文法に則り文の構文構造を求める構文解析手段および形態素解析手段によって切り出された単位文字列および／または単位文字列が統合されてなる単位文字列統合体の意味を解析する意味解析手段からなるものが用いられる。この場合、要約文生成手段としては、形態素解析手段、構文解析手段および意味解析手段の解析結果に基づいて、単位文字列および／または単位文字列統合体のうち、重要度の高いものを組み合わせて要約文を生成するものが用いられる。

【0025】この発明による第6の要約音声作成装置は、少なくとも音声が記録された記録媒体を、標準再生速度より高速で再生する手段、高速再生された音声を認識して、入力音声を文字コード列に変換する音声認識手段、文字コード列で表される文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成する手段、ならびに生成された要約文の文字コード列に応じた、高速再生音声の発声速度より遅い発声速度の音声を生成して出力する音声合成手段を備えていることを特徴とする。

【0026】この発明による映像・音声処理装置は、映像と音声とが対応づけられて記録された記録媒体を、標準再生速度より高速で再生する手段、高速再生された音声を認識して、入力音声を文字コード列に変換する音声認識手段、文字コード列で表される文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を生成する手段、生成された要約文の文字コード列に応じた、高速再生音声の発声速度より遅い発声速度の音声を生成する音声合成手段、ならびに高速再生された映像と音声合成手段によって生成された要約文に対する音声とを出力する出力手段を備えていることを特徴とする。

【0027】

【作用】この発明による第1の要約文作成装置では、まず、文の構成要素の重要度が判定される。そして、重要度の高い構成要素が組み合わされて要約文が生成される。

【0028】この発明による第2の要約文作成装置では、まず、言語解析手段によって、文が構成要素に分解されて文が解析される。そして、言語解析手段による解析結果に基づいて、重要度の高い構成要素が組み合わされて要約文が生成される。

【0029】この発明による第3の要約文作成装置では、まず、音声認識手段により、入力音声が認識され、入力音声が文字コード列に変換される。そして、文字コード列で表される文の構成要素の重要度が判定され、重要度の高い構成要素が組み合わされて要約文が生成される。

【0030】この発明による第1の要約音声作成装置では、文の構成要素の重要度が判定され、重要度の高い構成要素が組み合わされて要約文が生成される。そして、生成された要約文の文字コード列に応じた音声が、音声合成手段によって生成される。

【0031】この発明による第2の要約音声作成装置では、まず、音声認識手段により、入力音声が認識され、入力音声が文字コード列に変換される。次に、文字コード列で表される文の構成要素の重要度が判定され、重要度の高い構成要素が組み合わされて要約文が生成される。そして、生成された要約文の文字コード列に応じた音声が、音声合成手段によって生成される。

【0032】この発明による第3の要約音声作成装置では、音声認識手段により、入力音声が認識され、入力音声が文字コード列に変換される。次に、言語解析手段によって、文字コード列で表される文が構成要素に分解されて文が解析される。次に、言語解析手段による解析結果に基づいて、重要度の高い構成要素が組み合わされて要約文が生成される。そして、生成された要約文の文字コード列に応じた音声が、音声合成手段によって生成される。

【0033】この発明による第4の要約音声作成装置では、入力音声は記憶手段に記憶される。また、音声認識手段により、入力音声が認識され、入力音声が文字コード列に変換される。次に、文字コード列で表される文の構成要素の重要度が判定され、重要度の高い構成要素が組み合わされて要約文が生成される。そして、生成された要約文を構成する各構成要素に対応する単位音声が記憶手段から読み出されて、要約文に応じた音声が編集される。

【0034】この発明による第5の要約音声作成装置では、入力音声は記憶手段に記憶される。また、音声認識手段により、入力音声が認識され、入力音声が文字コード列に変換される。次に、言語解析手段によって、文字コード列で表される文が構成要素に分解されて文が解析される。次に、言語解析手段による解析結果に基づいて、重要度の高い構成要素が組み合わされて要約文が生

成される。そして、生成された要約文を構成する各構成要素に対応する単位音声が記憶手段から読み出されて、要約文に応じた音声が編集される。

【0035】この発明による第6の要約音声作成装置では、少なくとも音声が記録された記録媒体が、標準再生速度より高速で再生される。高速再生された音声が音声認識手段により認識され、入力音声が文字コード列に変換される。文字コード列で表される文の構成要素の重要度が判定され、重要度の高い構成要素が組み合わされて要約文が生成される。そして、生成された要約文の文字コード列に応じた、高速再生音声の発声速度より遅い発声速度の音声が生成されて出力される。

【0036】この発明による映像・音声処理装置では、映像と音声とが対応づけられて記録された記録媒体が、標準再生速度より高速で再生される。高速再生された音声が音声認識手段により認識され、入力音声が文字コード列に変換される。文字コード列で表される文の構成要素の重要度が判定され、重要度の高い構成要素が組み合わされて要約文が生成される。生成された要約文の文字コード列に応じた、高速再生音声の発声速度より遅い発声速度の音声が音声合成手段により生成される。そして、高速再生された映像と音声合成手段によって生成された要約文に対する音声とが出力される。

【0037】

【実施例】以下、図面を参照して、この発明の実施例について説明する。

【0038】図1は、要約文作成装置を示している。この要約文作成装置は、言語解析部1および要約文生成部2を備えている。言語解析部1には、文が文字コード列として入力される。言語解析部1は、文を単語、句等の構成要素に分解して文を解析する。要約文生成部2は、言語解析部1による解析結果に基づいて、重要度の高い構成要素を組み合わせて要約文を生成する。

【0039】図2、図3および図4は、それぞれ言語解析部1の具体例を示している。図2に示されている言語解析部1は、形態素解析辞書12を備えた形態素解析部11から構成されている。図3に示されている言語解析部1は、形態素解析辞書12を備えた形態素解析部11と、構文規則14を備えた構文解析部13とから構成されている。図4に示されている言語解析部1は、形態素解析辞書12を備えた形態素解析部11と、構文規則14を備えた構文解析部13と、意味辞書16を備えた意味解析部15とから構成されている。

【0040】形態素解析部11は、文を構成する単位文字列を切り出し、各単位文字列に関する文法情報を抽出する。単位文字列は通常、単語である。また、文法情報には、品詞、活用型等がある。なお、形態素解析処理は、たとえば、”講座　現在の言語7　「言語の機械処理」　長尾真　編　三省堂（1984年）”にそのアルゴリズムの一例が記載されているように、よく知られている技術である。

【0041】形態素解析辞書12は、ROM等の記憶手段からなる。図5は、形態素解析辞書12内に設けられた見出しテーブルの一部分の例を示している。この例では、「見出し」の単語ごとに、「品詞」、「付属情報」、「活用型」に関するデータが記憶されている。「見出し」の内容は、単語を示す文字コード（JISコード）列で表されている。

【0042】動詞、形容動詞に対応する単語の「活用型」の内容については、形態素解析辞書12内に設けられた活用テーブルに記憶されており、見出しテーブルの「活用型」には参照すべき活用テーブルを示すデータが記憶されている。活用テーブルの一例を図6に示しておく。

【0043】入力された文が、たとえば、「私たちの祖先は様々な工夫をしながら技術を蓄積してきた。」である場合には、形態素解析部11では、この文から「私たち」、「の」、「祖先」、「は」、「様々な」、「工夫」、「を」、「し」、「ながら」、「技術」、「を」、「蓄積」、「し」、「て」、「き」、「た」というように、単語が切り出され、各単語について品詞、付属情報が解析される。図7は、上記文例に対する形態素解析部11の解析結果の一例を示している。

【0044】言語解析部1が図2に示すように形態素解析部11によって構成されている場合には、要約文生成部2は、形態素解析部11の解析結果に基づいて、重要度の高い単語を組み合わせて要約文を生成する。この場合の重要度は、たとえば、次の表1に示すように、単語の品詞に応じて決定される。

【0045】

【表1】

品詞	重要度
名詞	高
代名詞	高
動詞	高
助詞	高
形容詞	中
形容動詞	中
副詞	低

【0046】表1に示すような重要度テーブルは、要約文生成部2が保持している。そして、要約文をどの重要度レベルまでの品詞を用いて生成するかは、たとえば、マニュアル操作によって設定される。重要度レベルとして「高」が指定された場合には、要約文生成部2は、重要度レベル「高」に対応する品詞である名詞、代名詞、動詞および助詞の単語を組み合わせて要約文を生成す

る。重要度レベルとして「中」が指定された場合には、要約文生成部２は、重要度レベル「中」と「高」に対応する品詞である名詞、代名詞、動詞、助詞、形容詞および形容動詞の単語を組み合わせて要約文を生成する。

【００４７】構文解析部１３は、形態素解析部１１の解析結果および構文規則１４に基づいて、文の構文構造を求める。なお、構文解析処理は、たとえば、”講座　現在の言語７　「言語の機械処理」　長尾真　編　三省堂（１９８４年）”にそのアルゴリズムの一例が記載されているように、よく知られている技術である。

【００４８】構文規則１４は、上記形態素解析辞書１２と同様に、ＲＯＭ等の記憶手段からなる。構文規則１４には、図８に一例が示されているように、品詞の結合状態と句との関係がテーブル形式で記憶されている。そして、形態素解析部１１によって得られた単語の品詞に基づいて、単語の統合が行われる。

【００４９】つまり、図８に示すような構文規則に基づいて、単語が統合されていく。図８の左側の構造が発見されれば、右側の句として定義される。また、統合の結果得られた名詞句は、図８の左側の名詞として更に統合されていく。

【００５０】たとえば、上記文例「私たちの…蓄積してきた。」においては、「私たち」＋「の」は、代名詞＋助詞の構造であるので、統合されて「私たちの」という名詞句として定義される。また、「祖先」＋「は」は、名詞＋助詞の構造であるので、統合されて「祖先は」という名詞句として定義される。さらに、このようにして得られた名詞句「私たちの」＋「祖先は」は、名詞＋名詞の構造であるので、統合されて「私たちの祖先は」という名詞句として定義される。

【００５１】上記文例「私たちの…蓄積してきた。」に対する統合結果が統合過程とともに図９に示されている。構文解析部１３は、統合過程において、統合によって得られた句の中に含まれる助詞、接続詞、動詞等の持つ属性、機能等によって、構文構造を解析していく。

【００５２】より具体的に説明すると、構文解析部１３は、統合過程において、次のような判断処理を行う。

【００５３】（１）まず、格助詞の存在によって、文要素の主格、目的格を検出する。また、動詞の存在によって、文要素の述部を検出する。上記文例「私たちの…蓄積してきた。」においては、格助詞「は」の存在により、名詞句「祖先は」を主格と判定し、格助詞「を」の存在により、名詞句「工夫を」と「技術を」とを目的格と判定し、動詞の存在により、動詞句「し」と「蓄積してきた」とを述部と判定する。

【００５４】（２）次に、上記（１）の判定結果に基づいて、主文「私たちの祖先は高度な技術を蓄積してきた」と、従文「私たちの祖先は様々な工夫をしながら」とから、文が構成されていると判定する。この例では、主文と従文とは、接続助詞「ながら」の存在によって判別されている。

【００５５】構文解析部１３の解析結果に基づいて最も短い要約文を作成した場合には、主文の主格と目的格と述部とからなる「祖先は技術を蓄積してきた」となる。

【００５６】言語解析部１が図３に示すように形態素解析部１１と構文解析部１３とによって構成されている場合には、要約文生成部２は、形態素解析部１１と構文解析部１３との解析結果に基づいて、重要度の高い句を組み合わせて要約文を生成する。

【００５７】句には、多数の分類があるが、主格、目的格および述部が文の骨格をなし、最も重要な句である。なお、その他の格には場所格、時間格等がある。句の重要度は、たとえば、次の表２のように決定される。

【００５８】

【表２】

句	重要度
主格	高
目的格	高
述部	高
その他の格	低

【００５９】表２に示すような重要度テーブルは、要約文生成部２が保持している。そして、要約部をどの重要度レベルまでの句を用いて生成するかは、たとえば、マニュアル操作によって設定される。また、主文と従文のうち、主文のみまたは両方を選択するかも、マニュアル操作によって設定される。

【００６０】意味解析部１５は、形態素解析部１１の解析結果および意味辞書１６に基づいて、形態素解析部１１によって切り出された単語、単語が統合された句等の意味を解析する。

【００６１】意味辞書１６は、上記形態素解析辞書１２と同様に、ＲＯＭ等の記憶手段からなる。意味辞書１６には、図１０に一例が示されているように、「見出し」の単語ごとに「意味情報」が記憶されている。この意味辞書１６によって、単語の意味情報が得られ、この単語を含む節や句の意味情報も検出することができる。すなわち、上記文例「私たちの…蓄積してきた。」中の単語「工夫」の意味情報が「科学技術」であるので、この単語を含む従文「私たちの祖先は様々な工夫をしながら」にも意味情報「科学技術」が割当られる。

【００６２】言語解析部１が図４に示すように形態素解析部１１と構文解析部１３と意味解析部１５とによって構成されている場合には、要約文生成部２は、形態素解析部１１と構文解析部１３と意味解析部１５との解析結果に基づいて、重要度の高い句を組み合わせて要約文を生成する。

【００６３】たとえば、マニュアル操作により、重要度

の高い意味情報として「科学技術」が指定されている場合には、この意味情報に該当する従文「私たちの祖先は様々な工夫をしながら」も主文同様に重要度が高いと判別される。したがって、この場合には、主文と従文のそれぞれの主格と目的格と述部とが組み合わされて、「祖先は工夫をしながら技術を蓄積してきた」という要約文が作成される。

【0064】なお、「意味情報」には、科学技術の他、政治経済、医学、法律等があり、これらの意味情報が重要度の高い意味情報として指定されることによって、指定された意味情報に対応する分野に焦点をあてた要約文が作成される。

【0065】上記各実施例では、単語または句の重要度レベルは、マニュアル操作等によって設定されているが、設定された要約率に応じて重要度レベルまたは従文の重要度等を自動的に決定するようにしてもよい。

【0066】ここで、要約率とは、原文の文章の長さに対する、要約文の文章の長さの比率をいう。言語解析部1が図3に示すように形態素解析部11と構文解析部13とから構成されている場合には、重要度レベルと従文の重要度との組み合わせによって次の表3のように4つの要約レベルの選択肢ができる。

【0067】

【表3】

	主文／従文	句（格）の高低
1	主文のみ	高
2	主文のみ	低
3	従文を含める	高
4	従文を含める	低

【0068】要約率が、たとえば1／3に設定されているとすると、上記4つの要約レベルの選択肢のうち、要約率が1／3に最も近くなるものを選択して要約文が作成される。

【0069】また、言語解析部1が図4に示すように形態素解析部11および構文解析部13の他、意味解析部15を含んでいるものでは、指定された重要な意味情報に基づいて、重要度の高い単語、文節、句の存在を加味して、上記表3の4つの要約レベルの選択肢のうちから、結果として要約率が1／3に最も近くなるものを選択して、要約文が作成される。この選択は、全ての選択肢について要約処理を行うことによって自動的に行われる。

【0070】図11は、要約音声作成装置を示している。この要約音声作成装置は、音声認識部21、要約文作成装置22、音声合成部23を備えている。

【0071】音声認識部21には、音声が入力される。音声認識部21は、入力音声を認識して、入力音声を文字コード列に変換する。つまり、音声認識部21は、単音節（かな1文字に相当）認識を行うものであり、予め記憶されている単音節ごとの参照音声パターンと、入力音声のパターンとが比較される。この比較処理には、連続ダイナミックプログラミングの手法などが使用される。このような比較処理の結果、入力音声のパターンに類似する参照音声パターンに対応した文字コード列が類似度とともに出力される。音声認識部21によって得られた文字コード列は、要約文作成装置22に送られる。

【0072】要約文作成装置22は、入力された文字コード列で表される文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を作成する。要約文作成装置22としては、上述した図2、図3または図4に示されている要約文作成装置1が用いられる。要約文作成装置22で作成された要約文は、文字コード列として音声合成部23に送られる。

【0073】音声合成部23は、要約文作成装置22で作成された要約文を、音声標準パターンを用いて、音声に変換して出力する。音声合成部23は、規則合成装置から構成されている。規則合成装置は、かな1文字を単位とするかまたは子音・母音・子音の組み合わせ（2000程度）を単位とした音声のセグメント信号波形を記憶した音素メモリを備えており、入力される文字コード列に対応するセグメント信号波形を接続する。この接続処理の際には、信号波形には、文字コードの配列に対応した規則に基づいて、代表的なアクセントや抑揚（基本周波数変化）が付加される。

【0074】図12は、他の要約音声作成装置を示している。この要約音声作成装置は、音声認識部21、要約文作成装置22、音声合成部23および声質変換部24を備えている。音声認識部21、要約文作成装置22および音声合成部23は、図11に示すものと同じであるので、その説明を省略する。

【0075】声質変換部24は、入力音声（要約音声作成装置に入力された音声）の音程および音質（音声スペクトル）に基づいて、音声合成部23で生成された音声の声質を、入力音声の声質に応じた音質に変換する。したがって、要約文の出力音声の声質が要約音声作成装置に入力された音声の声質に近似したものとなる。したがって、たとえば、入力音声が女性の声の場合には、出力音声も女性の声となる。また、入力音声が老人の声の場合には、出力音声も老人の声となる。つまり、入力音声の性別、年齢等に応じた出力音声が得られる。

【0076】図13は、さらに他の要約音声作成装置を示している。この要約音声作成装置は、バッファメモリ31、音声認識部32、要約文作成装置33および音声編集部34を備えている。音声認識部32および要約文作成装置33は、図11に示す音声認識部21および要約文作成装置22とそれぞれ同じである。

【0077】入力音声は、バッファメモリ31に蓄積さ

れ、音声認識部３２に順次送られる。音声認識部３２は、バッファメモリ３１から送られてきた入力音声を認識して、認識結果を文字コード列に変換して出力する。この際、認識結果に対応する音声のバッファメモリ３１内のアドレスも合わせて出力される。

【００７８】要約文作成装置３３は、音声認識部３２から入力された文字コード列で表される文の構成要素の重要度を判定し、重要度の高い構成要素を組み合わせて要約文を作成する。そして、重要度が高いと判定された構成要素に対応するアドレスとともに、生成された要約文が音声編集部３４に送られる。

【００７９】音声編集部３４は、入力されたアドレスに基づいて、要約文を構成する各構成要素に対応する単位音声をバッファメモリ３１から読み出して接続し、要約文に応じた音声波形を生成する。したがって、バッファメモリ３１に記憶されている入力音声中の重要な単位音声を繋ぎ合わせて要約文に対する音声（要約音声）が作成されるので、要約文に対する音声の声質は、入力音声の声質とほぼ同じになる。

【００８０】図１４は、さらに他の要約音声作成装置を示している。この要約音声作成装置は、バッファメモリ３１、音声認識部３２、要約文作成装置３３、音声編集部３４および韻律調整部３５を備えている。バッファメモリ３１、音声認識部３２および要約文作成装置３３は、図１３に示すものと同じであるので、その説明を省略する。

【００８１】音声編集部３４は、入力されたアドレスに基づいて、要約文を構成する各構成要素に対応する単位音声をバッファメモリ３１から読み出して接続し、要約文に応じた音声波形を生成する。生成された音声波形は、韻律調整部３５に送られる。この際、接続された各単位音声の継続時間長も、付帯情報として韻律調整部３５に送られる。

【００８２】韻律調整部３５は、音声編集部３４で編集された音声を構成する各単位音声の繋ぎ目を、アクセント調整等を行うことによって、なめらかにするものである。韻律調整部３５には、音声編集部３４からの要約音声波形および各単位音声の継続時間長の他、要約文生成装置３３からの要約文を表す文字コード列が送られる。

【００８３】図１５は、韻律調整部３５の構成を示している。アクセント生成部４１は、アクセント辞書４２およびアクセント変形規則４３を用いて、要約文作成装置３３から送られてきた要約文からアクセント情報を抽出する。たとえば、「そせん」であれば、「そ」の部分にアクセントがある。なお、このアクセント情報の抽出処理には、形態素解析と構文解析とが必要となるので、要約文作成装置３３の形態素解析部１１と構文解析部１３との両解析結果を利用することもできる。抽出されたアクセント情報は、ピッチパターン生成部４４に送られる。

【００８４】ピッチパターン生成部４４は、音声編集部３４から送られてきた各単位音声の継続時間長に基づいて、アクセントのあるところが高くなるようにピッチパターンを生成する。この生成方法としては、「藤崎モデル」に代表される手法がある。

【００８５】ピッチ抽出部４５では、音声編集部３４から送られてきた要約文の音声波形の実際のピッチパターンを抽出する。ピッチ抽出方法には、自己相関に基づく手法等、様々な手法が知られている。

【００８６】音程変換部４６は、ピッチ抽出部４５で抽出されたピッチパターンがピッチパターン生成部４４によって生成されたピッチパターンとなるように、要約文の音声波形の音程を変換して出力する。音程変換技術は、カラオケ装置等において、実用化されている技術である。

【００８７】次に、上述した要約文作成装置または要約音声作成装置の応用例について、説明する。

【００８８】図１６は、要約文作成装置を応用したディクテーションシステムを示している。

【００８９】マイク１０１から音声が入力された音声信号またはテープ再生部１０２によって再生された音声信号は、Ａ／Ｄ変換器１０３を介して音声認識部１１２に入力される。

【００９０】音声認識部１１２に入力された音声信号に対しては、音声認識部１１２とワードプロセッサ機能を有する文書処理部１１１とによって、音声入力ワープロ処理が行われる。この処理結果である入力音声に対する文字コード列からなる文章は、図示しないメインメモリ（ＲＡＭ）、フロービーディスク１１５、ハードディスク１１６等の記憶手段に記憶され、また、必要に応じてディスプレイ１１７に表示される。

【００９１】記憶手段に記憶された文章に対して、要約文作成装置１１３が要約文を自動的に作成する。作成された要約文は、メインメモリに記憶され、必要に応じてディスプレイ１１７で表示されたり、プリンタ１０６でプリントアウトされたりする。

【００９２】また、必要であれば、要約文作成装置１１３で作成された要約文は、音声合成部１１４によって音声信号に変換された後、Ｄ／Ａ変換器１０４を介してスピーカ１０５に送られて音声出力される。

【００９３】また、次のようにして、要約文を所定枚数の印刷用紙に納まるように作成することもできる。すなわち、文書処理部１１１に、例えばＡ４サイズの用紙１枚に要約書を作成する旨の指令を入力する。文書処理部１１１は、要約文作成装置１１３で設定可能な要約率のうち、Ａ４サイズの用紙１枚に記述できる範囲（文字数）内に要約文が納まる要約率を要約文作成装置１１３に設定する。このような設定を行うためのパラメータとしては、用紙サイズ、印刷文字のポイント数、用紙枚数等があり、これらのパラメータを指定することによっ

て、要約文の文字数が決定される。このような機能を用いれば、要約文からなる１枚の議事録を作成することができる。

【００９４】また、図１６のシステムにＯＣＲ（文字認識機能付）を入力手段として用いることができる。この場合には、多数枚の会議議事録をＯＣＲに認識させ、この認識結果に基づいて、要約文からなる、たとえば１枚の議事録を作成することができる。

【００９５】図１７は、要約音声作成装置をＶＴＲの高速再生時に適用した例を示している。

【００９６】キャプスタンサーボ回路２０１は、コントロールヘッド２０２からのコントロール信号およびキャプスタン２０３からの速度信号に基づいて、ビデオテープ２０５の走行速度が一定速度になるように、キャプスタンモータ２０４を制御する。なお、たとえば、２倍速再生時には、ビデオテープ２０５の走行速度が標準再生時の速度の２倍となるように、キャプスタンモータ２０４が制御される。

【００９７】ビデオヘッド２０６は、ビデオテープ２０５の映像トラックを再生する。ビデオヘッド２０６は、ヘッドスイッチング回路２０７により所定の順序で切り換え出力され、映像再生回路２０８で映像信号に変換される。

【００９８】オーデオヘッド２０９は、ビデオテープ２０５のオーディオトラックを再生する。再生された音声信号は、要約音声作成装置２００に送られる。

【００９９】高速再生時においては、ビデオテープに記録された映像と音声とが共に高速再生される。高速再生された映像は、モニタに表示される。高速再生された音声は、要約音声作成装置２００に送られる。要約音声作成装置２００は、高速再生音声の発声速度より遅い発声速度の要約音声を生成して出力する。たとえば、２倍速再生時には、標準再生速度の発声速度の要約音声が生成されて出力される。

【０１００】このように要約音声作成装置２００を、ＶＴＲの高速再生時に適用した場合には、高速再生音声の発声速度より遅い発声速度の要約音声を出力できるので、高速再生時の出力音声が聞き取り易くなる。また、要約音声が出力されているので、元の音声よりも言葉数が大幅に減少するため、要約音声が途切れるといったことも起こりにくくなる。なお、入力音声から要約音声を生成するための処理時間によって、出力映像と出力音声との間の時間ずれが目立つ場合には、画像メモリを用いるなどして、映像出力を遅延させ、出力映像と出力音声との同期を確保することも可能である。

【０１０１】要約音声作成装置の応用例としては、ＶＴＲの他、テープレコーダ、留守番電話機等があり、音声の簡略化と録音テープの節約化が図れる。

【０１０２】要約音声作成装置の留守番電話機への応用について説明する。留守番電話機は、外出中にかかってきた相手方のメッセージを録音する機能を備えており、外出中に電話をかけてきた相手方のメッセージ等を録音することができる。録音されているメッセージを聞くには、次のような方法がある。第１の方法は、帰宅した後、留守番電話機を操作して録音されているメッセージを再生して聞く。第２の方法は、外出先の電話から留守番電話機をリモート操作して、録音されているメッセージを再生して聞く。

【０１０３】このような留守番電話機を、たとえば、図１６のシステムのＡ／Ｄ変換器１０３およびＤ／Ａ変換器１０４に接続する。そして、留守番電話機内の録音メッセージ再生装置によって再生された音声信号をＡ／Ｄ変換器１０３に入力する。図１６のシステムにより、Ａ／Ｄ変換器１０３に入力された音声信号から要約音声が作成され、その音声信号がＤ／Ａ変換器１０４から出力される。

【０１０４】上記第１の方法で録音メッセージを聞く場合には、Ｄ／Ａ変換器１０４から出力された要約音声信号が留守番電話機から出力される。上記第２の方法で録音メッセージを聞く場合には、Ｄ／Ａ変換器１０４から出力された要約音声信号は、電話回線を介して外出先電話機に送られ、外出先電話機から出力される。いずれの方法においても、電話機から録音メッセージの要約音声が出力されるので、短時間にメッセージ内容を取得できる。また、上記第２方法の場合には、電話回線使用料の低廉化も図れる。

【０１０５】

【発明の効果】この発明によれば、文から自動的に要約文を作成することができる。また、この発明によれば、入力された音声から、自動的に要約文を作成することができる。また、この発明によれば、文から自動的に要約文を作成し、作成した要約文に対応する音声を出力することができる。また、この発明によれば、入力された音声から、自動的に要約文を作成し、作成した要約文に対応する音声を出力することができる。また、この発明によれば、ＶＴＲ等の再生装置の高速再生時において、高速再生された音声から標準速度の要約音声を作成して出力することができる。

【図面の簡単な説明】

【図１】要約文作成装置の構成を示すブロック図である。

【図２】言語解析部の構成を示すブロック図である。

【図３】言語解析部の他の例を示すブロック図である。

【図４】言語解析部のさらに他の例を示すブロック図である。

【図５】形態素解析辞書内の見出しテーブルの一例を示す模式図である。

【図６】形態素解析辞書内の活用テーブルの一例を示す模式図である。

【図７】形態素解析結果を示す模式図である。

【図８】構文規則の一例を示す模式図である。

【図９】構文解析結果およびその過程の一例を示す模式図である。

【図１０】意味辞書の内容の一例を示す模式図である。

【図１１】要約音声作成装置の構成を示すブロック図である。

【図１２】要約音声作成装置の他の例を示すブロック図である。

【図１３】要約音声作成装置のさらに他の例を示すブロック図である。

【図１４】要約音声作成装置のさらに他の例を示すブロック図である。

【図１５】韻律調整部の構成を示すブロック図である。

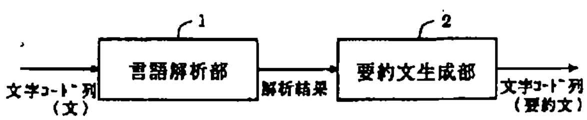
【図１６】要約書作成装置を適用したディクテーションシステムを示すブロック図である。

【図１７】要約書作成装置をＶＴＲに適用した応用例を示すブロック図である。　＊

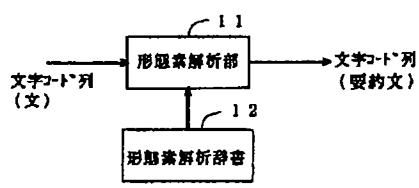
＊【符号の説明】

１　言語解析部
２　要約文生成部
１１　形態素解析部
１２　形態素解析辞書
１３　構文解析部
１４　構文規則
１５　意味解析部
１６　意味辞書
２１、３２　音声認識部
２２、３３、１１３　要約文作成装置
２３　音声合成部
２４　声質変換部
３１　バッファメモリ
３４　音声編集部
３５　韻律調整部
２００　要約音声作成装置

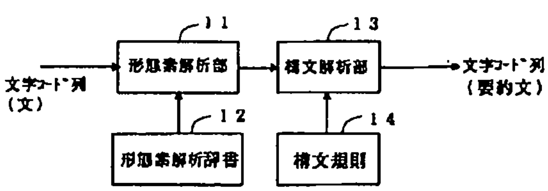
【図１】



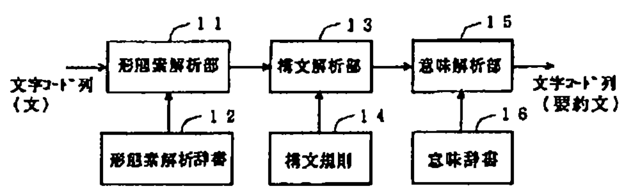
【図２】



【図３】



【図４】



【図６】

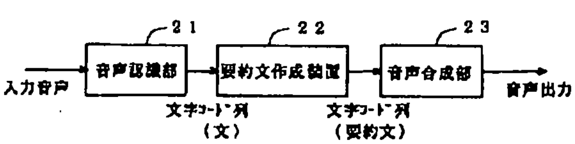
活用ﾃｰﾌﾞﾙの例（カ行変格活用ﾃｰﾌﾞﾙの場合）

未然形	連用形	終止形	連体形	仮定形	命令形
こ	き	くる	くる	くれ	こい

【図８】

名詞＋助詞 → 名詞句
形容詞＋名詞 → 名詞句
名詞＋名詞 → 名詞句
動詞＋助動詞 → 動詞句

【図１１】



【図5】

見出しテーブルの内容の例（一部分）

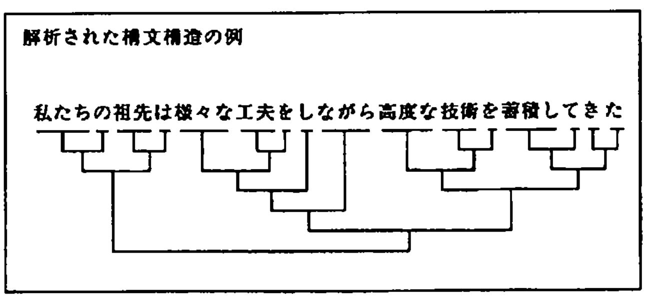
見出し	品詞	付属情報	活用型（参照テーブル）
くる	動詞	補助動詞	カ行変格活用テーブル
する	動詞		サ行変格活用テーブル
た	助動詞		助動詞テーブル
て	助詞	接続助詞	
ながら	助詞	接続助詞	
の	助詞	格助詞	
は	助詞	係助詞	
を	助詞	格助詞	
技術	名詞		
工夫	名詞		
高度だ	形容動詞		形容動詞テーブル
私たち	代名詞	一人称複数	
祖先	名詞		
蓄積する	動詞		サ行変格活用テーブル
様々だ	形容動詞		形容動詞テーブル

【図7】

形態素解析の結果

単語	品詞	付属情報
私たち	代名詞	一人称複数
の	助詞	格助詞
祖先	名詞	
は	助詞	係助詞
様々な	形容動詞	連体形
工夫	名詞	
を	助詞	格助詞
し	動詞	サ変・連用形
ながら	助詞	接続助詞
高度な	形容動詞	連体形
技術	名詞	
を	助詞	格助詞
蓄積し	動詞	サ変・連用形
て	助詞	接続助詞
き	動詞	補助動詞・カ変・連用形
た	助動詞	終止形

【図9】

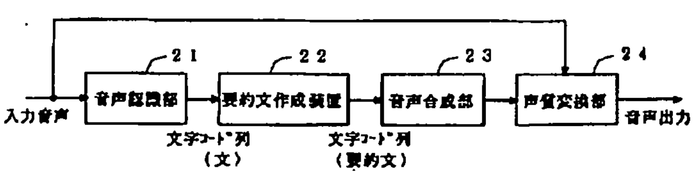


【図10】

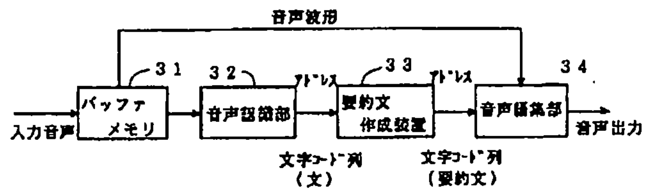
意味辞書の内容の例（一部分）

見出し	意味情報
くる	・・・
する	・・・
た	・・・
て	・・・
ながら	・・・
の	・・・
は	・・・
を	・・・
技術	科学技術
工夫	科学技術
高度だ	程度
私たち	人
祖先	生物
蓄積する	・・・
様々だ	・・・

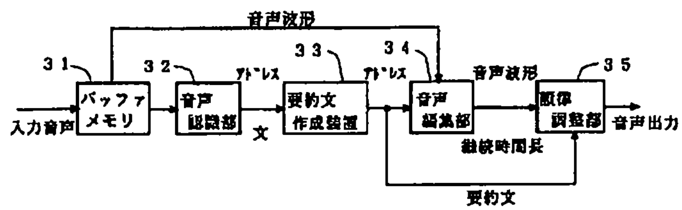
【図12】



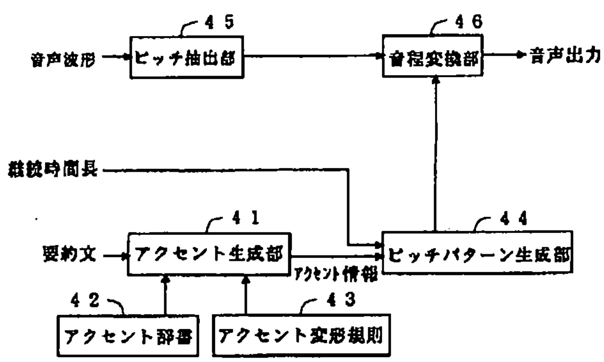
【図１３】



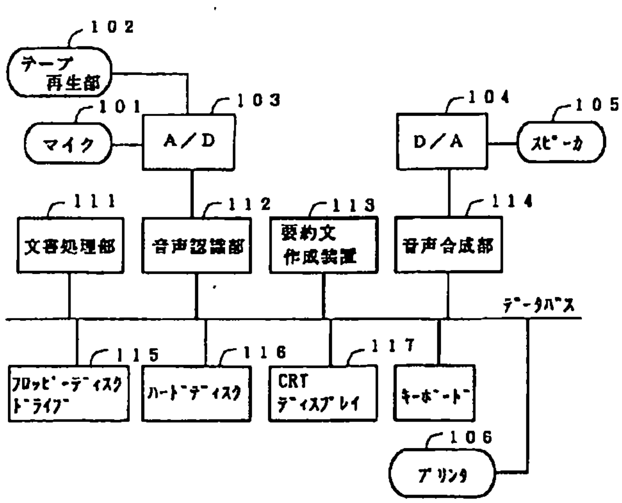
【図１４】



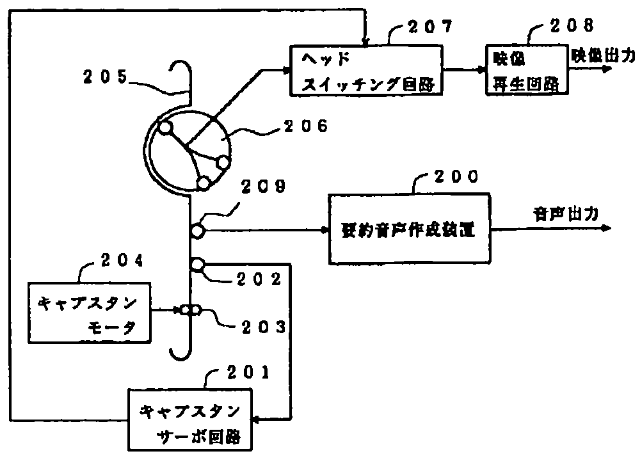
【図１５】



【図１６】



【図17】


207
ヘッド
スイッチング回路
208
映像
再生回路
映像出力
205
206
209
200
要約音声作成装置
音声出力
204
キャプスタン
モータ
202
203
201
キャプスタン
サーボ回路

Tab C

OIPE IAP35
AUG 06 2007
PATENT & TRADEMARK OFFICE

CERTIFICATE OF MAILING UNDER 37 C.F.R. §1.8

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail, with sufficient postage, in an envelope addressed to: Commissioner for Patents, P. O. Box 1450, Alexandria, VA 22313-1450, on the below date:

Date: August 2, 2007 Name: Tadashi Horie (Reg. No. 40,437) Signature: Tadashi Horie

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Appln. of: Toshihiko Oba
Appln. No.: 09/673,360
Filed: October 16, 2000
For: SPEECH TRANSFORMATION METHOD AND APPARATUS
Attorney Docket No: 11934/3

Examiner: Pierre Myriam
Art Unit: 2626

SIXTH SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT

Mail Stop Amendment
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

In accordance with the duty of disclosure under 37 C.F.R. §1.56 and §§1.97-1.98, and more particularly in accordance with 37 C.F.R. §1.97(c), Applicant hereby cites the following reference(s):

U.S. PATENT DOCUMENTS

DOCUMENT NUMBER Number-kind Code (if known)	DATE	NAME
5,839,109	November 17, 1998	Iwamida
5,326,349	July 5, 1994	Baraff
5,283,833	February 1, 1994	Curhch
4,425,481	January 10, 1984	Mansgold et al.

BRINKS HOFER GILSON &LIONE 371667.1

FOREIGN DOCUMENTS

DOCUMENT NUMBER Number-Kind Code (if known)	DATE	COUNTRY
GB 2 256 959 A	December 23, 1992	GB
WO 97/29482 A	August 14, 1997	WIPO
EP 0 872 808 A1	October 21, 1998	EPO

OTHER DOCUMENTS

Supplementary European Search Report issued May 4, 2007 in the European Application No. 00903984.3

Applicant is enclosing Form PTO-1449 (one sheet), along with a copy of each listed reference for which a copy is required under 37 C.F.R. §1.98(a)(2). As each of the listed references is in English, no further commentary is believed to be necessary, 37 C.F.R §1.98(a)(3). Applicant respectfully requests the Examiner's consideration of the above reference(s) and entry thereof into the record of this application.

Reference G8 is a Supplementary European Search Report issued in the counterpart European Application. References G1 to G7 were cited in the Search Report.

By submitting this Statement, Applicant is attempting to fully comply with the duty of candor and good faith mandated by 37 C.F.R. §1.56. As such, this Statement is not intended to constitute an admission that any of the enclosed references, or other information referred to therein, constitutes "prior art" or is otherwise "material to patentability," as that phrase is defined in 37 C.F.R. §1.56(a).

Applicant certifies under 37 C.F.R. §1.97(e)(1) that each item of information in this Statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of this Statement (a copy of any foreign communication first citing a listed reference is attached for the Examiner's reference). Accordingly, Applicant has calculated no fee to

be due in connection with the filing of this Statement. However, the Director is authorized to charge any fee deficiency associated with the filing of this Statement to a deposit account, as authorized in the Transmittal accompanying this Statement.

Respectfully submitted,

August 2, 2007
Date

Tadashi Horie (Reg. No. 40,437)

OIPE
AUG 0 6 2007
IAP35
PATENT & TRADEMARK OFFICE

FORM PTO-1449	SERIAL NO. 09/673,360	CASE NO. 11934/3
LIST OF PATENTS AND PUBLICATIONS FOR APPLICANT'S INFORMATION DISCLOSURE STATEMENT	FILING DATE October 16, 2000	GROUP ART UNIT 2626
(use several sheets if necessary)	APPLICANT(S): Toshihiko Oba	

REFERENCE DESIGNATION **U.S. PATENT DOCUMENTS**

EXAMINER INITIAL		DOCUMENT NUMBER Number-Kind Code (if known)	DATE	NAME	CLASS/ SUBCLASS	FILING DATE
	G1	5,839,109	11/17/1998	Iwamida		
	G2	5,326,349	07/05/1994	Baraff		
	G3	5,283,833	02/01/1994	Curhch		
	G4	4,425,481	01/10/1984	Mansgold et al.		

FOREIGN PATENT DOCUMENTS

EXAMINER INITIAL		DOCUMENT NUMBER Number-Kind Code (if known)	DATE	COUNTRY	CLASS/ SUBCLASS	TRANSLATION YES OR NO
	G5	GB 2 256 959 A	12/23/1992	GB		
	G6	WO 97/29482 A	08/14/1997	WIPO		
	G7	EP 0 872 808 A1	10/21/1998	EPO		

EXAMINER INITIAL	OTHER ART – NON PATENT LITERATURE DOCUMENTS (Include name of author, title of the article (when appropriate), title of the item (book, magazine, journal, serial, symposium, catalog, etc.), date page(s), volume-issue number(s), publisher, city and/or country where published.	
	G8	Supplementary European Search Report issued May 4, 2007 in the European Application No. 00903984.3

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant.

(12) UK Patent Application (19) GB (11) 2 256 959 (13) A

(43) Date of A publication 23.12.1992

(21) Application No 9113466.8

(22) Date of filing 21.06.1991

(71) Applicant
Nigel Glyn Wallace
33 West Hill Road, Foxton, Cambridgeshire, CB2 6SZ, United Kingdom

(72) Inventor
Nigel Glyn Wallace

(74) Agent and/or Address for Service
Nigel Glyn Wallace
33 West Hill Road, Foxton, Cambridgeshire, CB2 6SZ, United Kingdom

(51) INT CL^5
G10L 3/00

(52) UK CL (Edition K)
G4R REX RRL R1F R1X R10A R11D R11E R9B R9C
U1S S1827 S2409

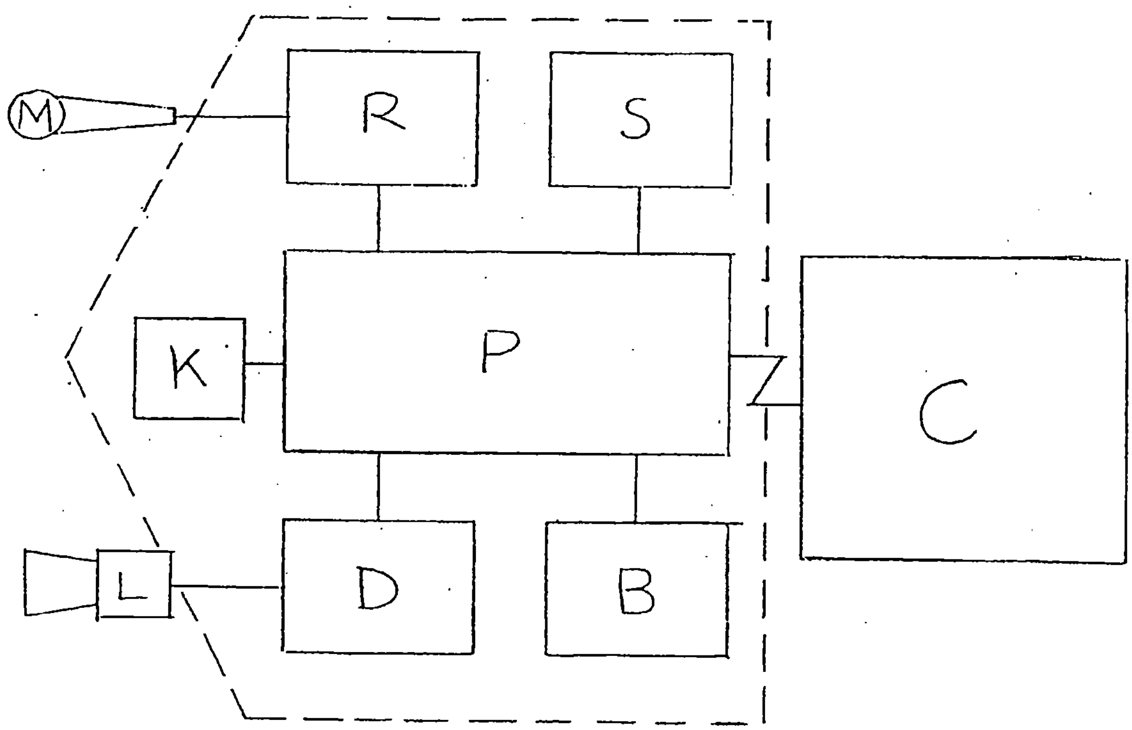
(56) Documents cited
None

(58) Field of search
UK CL (Edition K) G4R REX RHA RHB RPC RPW RRL RRM RRP
INT CL^5 G10L
Online databases: WPI

(54) Speech recognition device

(57) An electronic device contained in a conveniently small box suitable for mounting on a wheelchair converts the inarticulate sounds a severely physically disabled and speed impaired person is able to make into normal intelligent conversation. It does this by allowing prerecorded words and phrases to be selected from a display screen by a code made up from a few simple sounds spoken into a microphone. The words and phrases are spoken through a loudspeaker also mounted on the wheelchair.

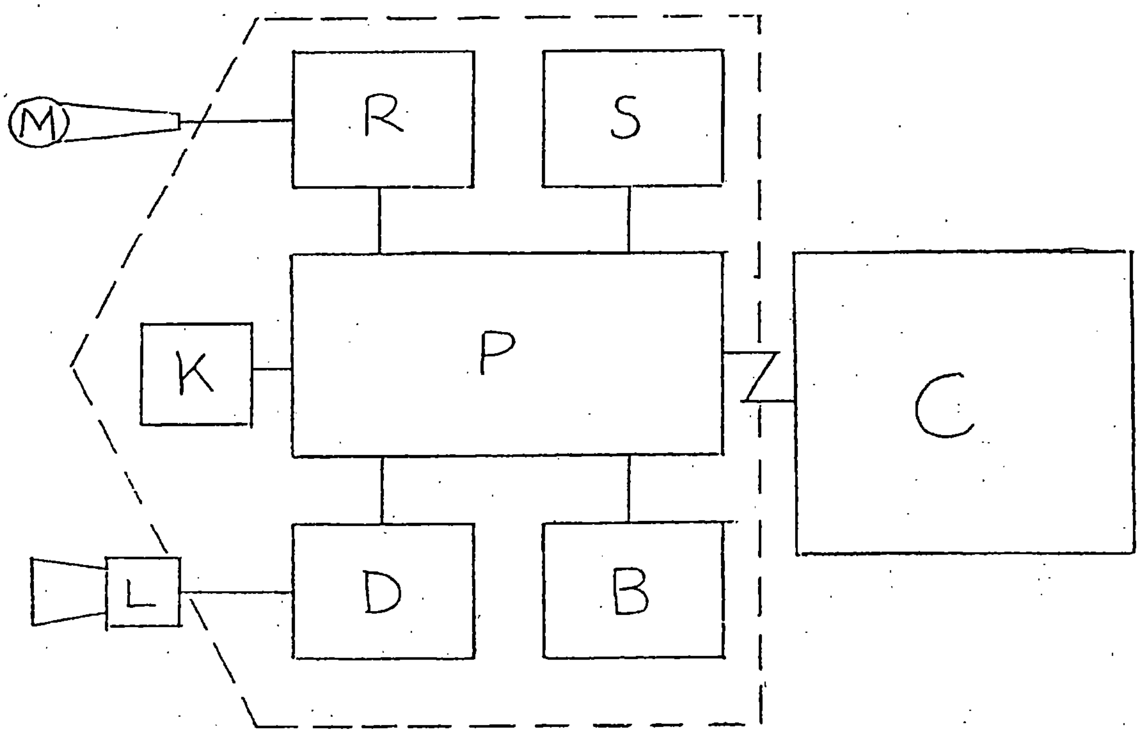
DIAGRAM OF GRUNT CONVERTER



GB 2 256 959 A

This print takes account of replacement documents submitted after the date of filing to enable the application to comply with the formal requirements of the Patents Rules 1990.

DIAGRAM OF GRUNT CONVERTER



PATENT SPECIFICATION

SPEECH RECOGNITION DEVICE

Device for assisting persons who can only make grunting noises to produce normal speech.

I, Nigel Glyn Wallace, a British Subject of 33 West Hill Road, Foxton, Cambridgeshire CB2 6SZ, do hereby declare the invention, for which I pray that a patent may be granted me, and the method by which it is to be performed, to be particularly described in and by the following statement :- This invention relates to a speaking device for physically disabled people of normal intelligence and hearing capability who are unable to enunciate words but may only emit a limited range of grunting sounds.

Speaking devices for such people are available by which keys or buttons are pressed to actuate speech producers, or text is entered by them from a typewriter type keyboard which is converted into synthesised speech. Many such people however are also highly spastic and severely physically disabled, such that they are unable to operate keys and buttons.

The object of this invention is to enable those who are capable of a limited range of sounds to convert them into intelligible speech.

According to this invention there is provided a combinaton of existing electronic devices which are controlled by a novel form of computer program instructions to enable the complete process to be performed without the aid of a normal able-bodied person or helper. A microphone is placed near to the disabled person's mouth or throat and a loudspeaker is placed at a suitable distance. The devices in between are contained in a convenient box which could be carried on a wheelchair and would be battery operated. The microphone signals are intercepted by a speech pattern recognition circuit which compares the incoming sounds with sound templates held in a Random Access Memory unit. The recognised sounds are used to select from a library of words and phrases held in the same memory unit. These coded words are turned into audio signals by a voice digitising circuit, the words and phrases having already been entered into the memory by a person capable of speech. Alternatively, a synthetic voice can be used with very much less memory requirement from a standard speech synthesiser integrated circuit.

The actions so far described are controlled by a microprocessor circuit which itself responds to grunt sounds. To enable the disabled person to monitor the operation and select the phrases required a standard liquid crystal screen is provided.

Such a device, in accordance with this invention, will now be described, by way of example only, with references to the accompanying diagram of the Grunt Converter.

The Grunt Converter is a device to convert the limited sounds available to a speech impaired person, such as a Dysarthria sufferer, into easily recognised speech. It consists of a microprocessor and memory unit (P) programmed to accept signals from a microphone (M) by way of a standard speech recogniser circuit (R), a program to store these signals in ASCII form in memory, a look-up table to find the words and phrases in the desired order displayed on a screen (S) and a standard speech digitiser (D) to transmit already recorded words and phrases to an amplifier and loudspeaker (L).

The Dysarthria sufferer/user needs to train the speech recogniser to recognise his/her grunts (this is done with a speech therapist in attendance). These sounds are described and entered on a chart on the display screen (S). Another person, male or female, selects the words and phrases most used by the disabled person and enters them into the memory of the Grunt Converter. Both these operations use the standard procedures prescribed by the makers of the units and are carried out using a standard PC computer (C). They are transferred to the Grunt Converter memory and held by battery (B) back-up permanently in the unit. The conversion programs are held on EPROM permanently and only limited controls are required to operate the unit by the Dysarthria sufferer. The phrases and words are accessed by up to 4 different grunt sounds. The user can by himself make up the phrases he or she needs from a large vocabulary of words and the phrases can be stored for future use. In cases where the user has difficulty recognising written words the graphic symbols representing words and phrases are displayed on the screen.

The Grunt Converter is contained in a small box with a window display similar to a PC portable (lap top) computer and a limited number of operating buttons (K). It is connected to a microphone (M) for grunt input and a loudspeaker (L) for speech output, and is designed to fit conveniently onto a wheelchair. It is battery operated with a mains adaptor for charging.

A further development of the Grunt Converter uses the principle of Context Selection such that screens of words and phrases are presented such that the words (or phrases) are those which naturally follow the previously selected word. By this means the number of grunts for selection is much reduced and at the same time the range of words available is greatly increased.

Another application for the Grunt Converter configuration is by way of a program for speech training. A speech therapist provides a series of spoken words of increasing difficulty which the speech impaired person tries to match. At each try the device assesses his/her accuracy and the speaking voice, pre-recorded by the therapist, encourages further attempts.

The described grunt converter device has the following advantages :-

1. Once set up it can be initiated and closed down by the disabled person without outside help.

2. It enables speech impaired persons to call up intelligible words and phrases without having to use physical movements.

3. It can be used by several such disabled persons together in a group whereby the individuals concerned can select the appropriate speaking voices for themselves.

4. If the speech recognition templates are entered by a speech therapist the device can be used to train a speech impaired person to improve his or her ennunciation.

The apparatus which has been described utilises electronic semiconducting devices of the type that are normally utilised in micro-computers.

WHAT I CLAIM IS :-

1. A speech recognition device which allows speech impaired persons to select intelligible words and phrases of their choice by uttering inarticulate sounds without any physical action on their part, using already existing circuits in a novel combination controlled by microprocessor machine instructions which themselves are initiated by the disabled users of the device.

2. A speech recognition device according to claim 1 wherein the words and phrases selected are projected audibly with a pre-recorded speaking voice to the disabled person's choice.

3. A speech recognition device according to claim 1 and 2 wherein the sound activated controls allow several different speech impaired persons to talk to each other with different speaking voices.

4. A speech recognition device according to claims 1 to 3 wherein the speech recognition templates are entered by a therapist for the speech impaired person to attempt to match and thus be trained to speak correctly.

5. A speech recognition device according to claims 1 to 4 wherein means are included to automatically assess and correct the speech impaired person under training.

6. A speech recognition device substantially as hereinfore described with reference to the accompanying diagram.

DESCRIPTION OF DRAWING.

M Microphone

R Pattern Recognition unit

P Microprocessor and memory unit

S Display Screen

D Speech Digitiser unit

L Loudspeaker

B Battery supply

C Setting up computer with keyboard

K Small keyboard for switching on by ablebodied person

Patents Act 1977
Examiner's report to the Comptroller under Section 17 (The Search Report)

Application number
GB 9113466.8

Relevant Technical fields

(i) UK Cl (Edition K) G4R (REX, RHA, RHB, RPC, RPW, RRL, RRM, RRP)

(ii) Int Cl (Edition 5) G10L

Databases (see over)

(i) UK Patent Office

(ii) ONLINE DATABASE: WPI

Search Examiner
J DONALDSON

Date of Search
21 AUGUST 1992

Documents considered relevant following a search in respect of claims 1 TO 6

Category (see over)	Identity of document and relevant passages	Relevant to claim(s)
	NONE	

Category	Identity of document and relevant passages	Relevant to claim(s)

Categories of documents

X: Document indicating lack of novelty or of inventive step.

Y: Document indicating lack of inventive step if combined with one or more other documents of the same category.

A: Document indicating technological background and/or state of the art.

P: Document published on or after the declared priority date but before the filing date of the present application.

E: Patent document published on or after, but with priority date earlier than, the filing date of the present application.

&: Member of the same patent family, corresponding document.

Databases: The UK Patent Office database comprises classified collections of GB, EP, WO and US patent specifications as outlined periodically in the Official Journal (Patents). The on-line databases considered for search are also listed periodically in the Official Journal (Patents).

Published 1992 at The Patent Office, Concept House, Cardiff Road, Newport, Gwent NP9 1RH. Further copies may be obtained from Sales Branch, Unit 6, Nine Mile Point, Cwmfelinfach, Cross Keys, Newport, NP1 7HZ. Printed by Multiplex techniques ltd. St Mary Cray, Kent.

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G10L 9/00

A1

(11) International Publication Number: WO 97/29482

(43) International Publication Date: 14 August 1997 (14.08.97)

(21) International Application Number: PCT/US97/01490

(22) International Filing Date: 28 January 1997 (28.01.97)

(30) Priority Data:
08/597,589 6 February 1996 (06.02.96) US

(71) Applicant: THE REGENTS OF THE UNIVERSITY OF CALIFORNIA [US/US]; 22nd floor, 300 Lakeside Drive, Oakland, CA 94612 (US).

(72) Inventors: HOLZRICHTER, John, F.; 200 Hillcrest Road, Berkeley, CA 94705 (US). NG, Lawrence, C.; 80 Country Hills Court, Danville, CA 94506 (US).

(74) Agent: SARTORIO, Henry, P.; P.O. Box 808, L-703, Livermore, CA 94550 (US).

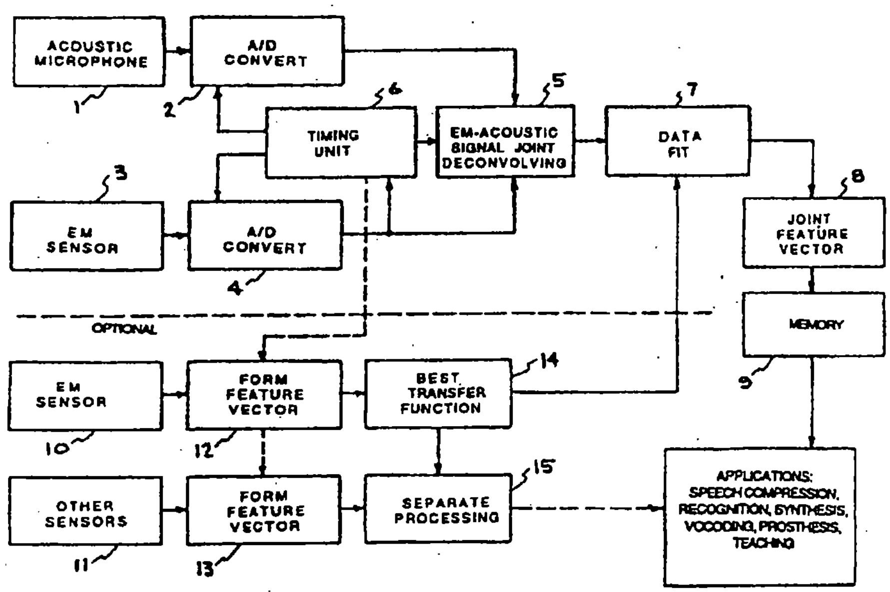
(81) Designated States: JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: SPEECH CODING, RECONSTRUCTION AND RECOGNITION USING ACOUSTICS AND ELECTROMAGNETIC WAVES

(57) Abstract

The use of EM radiation in conjunction with simultaneously recorded speech information enables a complete mathematical coding of acoustic speech. The methods include the forming of a feature vector (12, 13) for each pitch period of voiced speech and the forming of feature vectors (12, 13) for each time frame of unvoiced, as well as for combined voiced and unvoiced speech. The methods include how to deconvolve the speech excitation function from the acoustic speech output to describe the transfer function (7) each time frame. The formation of feature vectors (12, 13) defining all acoustic speech units over well-defined time frames can be used for purposes of speech coding, speech compression, speaker identification, language-of-speech identification, speech recognition, speech synthesis, speech translation, speech telephony, and speech teaching.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgystan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic	KR	Republic of Korea	SE	Sweden
CG	Congo	KZ	Kazakhstan	SG	Singapore
CH	Switzerland	LI	Liechtenstein	SI	Slovenia
CI	Côte d'Ivoire	LK	Sri Lanka	SK	Slovakia
CM	Cameroon	LR	Liberia	SN	Senegal
CN	China	LT	Lithuania	SZ	Swaziland
CS	Czechoslovakia	LU	Luxembourg	TD	Chad
CZ	Czech Republic	LV	Latvia	TG	Togo
DE	Germany	MC	Monaco	TJ	Tajikistan
DK	Denmark	MD	Republic of Moldova	TT	Trinidad and Tobago
EE	Estonia	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	UG	Uganda
FI	Finland	MN	Mongolia	US	United States of America
FR	France	MR	Mauritania	UZ	Uzbekistan
GA	Gabon			VN	Viet Nam

SPEECH CODING, RECONSTRUCTION AND RECOGNITION USING ACOUSTICS AND ELECTROMAGNETIC WAVES

The United States Government has rights in this invention pursuant to Contract No. W-7405-ENG-48 between the United States Department of Energy and the University of California for the operation of Lawrence Livermore National Laboratory.

BACKGROUND OF THE INVENTION

The invention relates generally to the characterization of human speech using combined EM wave information and acoustic information, for purposes of speech coding, speech recognition, speech synthesis, speaker identification, and related speech technologies.

Speech Characterization and Coding:

The history of speech characterization, coding, and generation has spanned the last one and one half centuries. Early mechanical speech generators relied upon using arrays of vibrating reeds and tubes of varying diameters and lengths to make human-voice-like sounds. The combinations of excitation sources (e.g., reeds) and acoustic tracts (e.g., tubes) were played like organs at theaters to mimic human voices. In the 20th century, the physical and mathematical descriptions of the acoustics of speech began to be studied intensively and these were used to enhance many commercial products such as those associated with telephony and wireless communications. As a result, the coding of human speech into electrical signals for the purposes of transmission was extensively developed, especially in the United States at the Bell Telephone Laboratories. A complete description of this early work is given by J. L. Flanagan, in "Speech Analysis, Synthesis, and Perception", Academic Press, N.Y., 1965. He describes the physics of speech and the mathematics of describing acoustic speech units (i.e., coding). He gives examples of how human vocal excitation sources and the human vocal tracts behave and interact with each other to produce human speech.

The commercial intent of the early telephone work was to understand how to use the minimum bandwidth possible for transmitting acceptable vocal quality on the then-limited number of telephone wires and on the limited frequency spectrum available for radio (i.e. wireless) communication. Secondly, workers learned that analog voice transmission uses typically 100 times more bandwidth than the transmission of the same word if simple numerical codes representing the speech units such as phonemes or words are transmitted. This technology is called "Analysis-Synthesis Telephony" or "Vocoding". For example, sampling at 8 kHz and using 16 bits per analog signal value requires 128 kbps, but the Analysis Synthesis approach can lower the coding requirements to below 1.0 kbps. In spite of the bandwidth advantages, vocoding has not been used widely because it requires accurate automated phoneme coding and resynthesis; otherwise the resulting speech tends to have a "machine accent" and be of limited intelligibility. One major aspect of the difficulty of speech coding is adequacy of the excitation information, including the pitch measurement, the voiced-unvoiced discrimination, and the spectrum of the glottal excitation pulse.

Progress in speech acoustical understanding and mathematical modeling of the vocal tract has continued and become quite sophisticated, mostly in the laboratory. It is now reasonably straightforward to simulate human speech by using differential equations which describe the increasingly complex concatenations of sound excitation sources, vocal tract tubes, and their constrictions and side branches (e.g., vocal resonators). Transform methods (e.g. electrical analogies solved by Fourier, Laplace, Z-transforms, etc.) are used for simpler cases and sophisticated computational modeling on supercomputers for increasingly complex and accurate simulations. See Flanagan (ibid.) for early descriptions of modeling, and Schroeter and Sondhi, "A hybrid time-frequency domain articulator speech synthesizer", IEEE Trans. on Acoustic Speech, ASSP 35(7) 1987 and "Techniques for Estimating Vocal-Tract Shapes from the Speech Signal", ASSP 2(1), 1343, 1994. These papers reemphasize that it is not possible to work backwards from the acoustic output to obtain a unique

mathematical description of the combined vocal fold--vocal tract system, which is called the "inverse problem" herein. It is not possible to obtain information that separately describes both the "zeros" in speech air flow caused by glottal (i.e., vocal fold) closure and those caused by closed, or resonant structures in the vocal tract. As a result, it is not possible to use the well developed mathematics of modern signal acquisition, processing, coding, and reconstructing to the extent needed.

In addition, given a mathematical vocal system model, it remains especially difficult to associate it with a unique individual because it is very difficult to obtain the detailed physiological vocal tract features of a given individual such as tract lengths, diameters, cross sectional shapes, wall compliance, sinus size, glottal size and compliance, lung air pressure, and other necessary parameters. In some cases, deconvolving the excitation source from the acoustic output can be done for certain sounds where the "zeros" are known to be absent, so the major resonant structures such as tract lengths can be determined. For example, simple acoustic resonator techniques (see the 1976 US patent 4,087,632 by Hafer) are used to derive the tongue body position by measuring the acoustic formant frequencies (i.e., the vocal tube resonance frequencies) and to constrain the tongue locations and tube lengths against an early, well known vocal tract model by Coker, "A Model of Articulatory Dynamics and Control", Proc. of IEEE, Vol.64(4), 452-460, 1976. The problem with this approach is that only gross dimensions of the tract are obtained, but detailed vocal tract features are needed to unambiguously define the physiology of the human doing the speaking. For more physiological details, x-ray imaging of the vocal tract has been used to obtain tube lengths, diameters, and resonator areas and structures. Also the optical laryngoscope, inserted into the throat, to view the vocal fold open and close cycles, is used in order to observe their sizes and time behavior.

The limit to further performance improvements in acoustic speech recognition, in speech synthesis, in speaker identification, and other related technologies is directly related to our inability to accurately solve the inverse problem. Present workers are unable to use acoustic speech output to work backwards to accurately

and easily determine the vocal tract transfer function, as well as the excitation amplitude versus time. The "missing" information about the separation of the excitation function from the vocal tract transfer function leads to many difficulties in automating the coding of the speech for each speech time frame and in forming speech sound-unit libraries for speech-related technologies. A major reason for the problem is that workers have been unable to measure the excitation function in real time. This has made it difficult to automatically identify the start and stop of each voiced speech segments over which a speech sound unit is constant. This has made it difficult to join (or to unjoin) the transitions between sequential vocalized speech units (e.g., syllables, phonemes or multiplets of phonemes) as an individual human speaker articulates sounds at rates of approximately 10 phonemes per second or two words per second.

The lack of precision in speech segment identification adds to the difficulty in obtaining accurate model coefficients for both the excitation function and the vocal tract. Further, this leads to inefficiencies in the algorithms and the computational procedures required by the technological application such as speech recognition. In addition, the difficulties described above prevent the accurate coding of the unique acoustic properties of a given individual for personalized, human speech synthesis or for pleasing vocoding. In addition, the "missing" information prevents complete separation of the excitation from the transfer function, and limits accurate speaker-independent speech-unit coding (speaker normalization). The incomplete normalization limits the ability to conduct accurate and rapid speech recognition and/or speaker identification using statistical codebook lookup techniques, because the variability of each speaker's articulation adds uncertainty in the matching process and requires additional statistical processing. The missing information and the timing difficulties also inhibit the accurate handling of co-articulation, incomplete articulation, and similar events where words are run together in the sequences of acoustic units comprising a speech segment.

In the 1970s, workers in the field of speech recognition showed that short "frames" (e.g., 10 ms intervals) of the time waveform

of a speech signal could be well approximated by an all poles (but no zeros) analytic representation, using numerical "linear predictive coding" (LPC) coefficients found by solving covariance equations. Specific procedures are described in B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Soc. Am. 50(2), pp. 63, 1971. The LPC coefficients are a form of speech coding and have the advantage of characterizing acoustic speech with a relatively small number of variables-- typically 20 to 30 per frame as implemented in today's systems. They make possible statistical table look up of large numbers of word representations using Hidden Markov techniques for speech recognition.

In speech synthesizers, code books of acoustic coefficients (e.g., using well known LPC, PARCOR, or similar coefficients) for each of the phonemes and for a sufficient number of diphonemes (i.e. phoneme pairs) are constructed. Upon demand from text-to-speech generators, they are retrieved and concatenated to generate synthetic speech. However, as an accurate coding technique, they only approximate the speech frames they represent. Their formation and use is not based upon using knowledge of the excitation function, and as a result they do not accurately describe the condition of the articulators. They are also inadequate for reproducing the characteristics of the given human speaker. They do not permit natural concatenation into high quality natural speech. They can not be easily related to an articulatory speech model to obtain speaker-specific physiological parameters. Their lack of association with the articulatory configuration makes it difficult to do speaker normalization, as well as to deal with the coarticulation and incomplete articulation problem of natural speech.

Present Example of Speech Coding:

Rabiner, in "Applications of Voice Processing to Telecommunications" Proc. of the IEEE 82, 199 Feb. 1994 points out that several modern text-to-speech synthesis systems in use today by AT&T use 2000 to 4000 diphonemes, which are needed to simulate the phoneme-to-phoneme transitions in the concatenation process for natural speech sounds. Figure 1 shows a prior art open loop acoustic speech coding system in which acoustic signals from a microphone are

processed, e.g. by LPC, and feature vectors are produced and stored in a library. Rabiner also points out (page 213) that in current synthesis models, the vocal source excitation and the vocal tract interaction "is grossly inadequate", and also that "when natural duration and pitch are copied onto a text-to-speech utterance, ... the quality of the ... synthetic speech improves dramatically." Presently, it is not possible to economically capture the natural pitch duration and voiced air-pulse amplitude vs. time, as well as individual vocal tract qualities, of a given individual's voice in any of the presently used models, except by very expensive and invasive laboratory measurements and computations.

J. L. Flanagan, "Technologies for Multimedia Communications", Proc. IEEE 82, 590, April 1994, describes low bandwidth speech coding: "At fewer than 1 bit per Nyquist sample, source coding is needed to additionally take into account the properties of the signal generator (such as voiced/unvoiced distinctions in speech, and pitch, intensity, and formant characteristics)." There is no presently, commercially useful method to account for the speech excitation source in order to minimize the coding complexity and subsequent bandwidth.

EM Sensors and Acoustic Information:

The use of EM sensors for measuring speech organ conditions for the purposes of speech recognition and related technologies are described in copending U.S. Patent Application, Ser. No. 08/597,596, filed 2/6/96, by Holzrichter. Although it has been recognized for many decades in the field of speech recognition that speech organ position and motion information could be useful, and EM sensors (e.g. rf and microwave radars) were available to do the measurement, no one had suggested a system using such sensors to detect the motions and locations of speech organs. Nor had anyone described how to use this information to code each speech unit and to use the code in an algorithm to identify the speech unit, or for other speech technology applications such as synthesis. Holzrichter showed how to use EM sensor information with simultaneously obtained acoustic data to obtain the positions of vocal organs, how to define feature vectors from this organ information to use as a coding technique, and how to use this information to do high-accuracy speech

recognition. He also pointed out that this information provided a natural method of defining changes in each phoneme by measuring changes in the vocal organ conditions, and he described a method to automatically define each speech time frame. He also showed that "photographic quality" EM wave images, obtained by tomographic or similar techniques, were not necessary for the implementation of the procedures he described, nor for the procedures described herein.

SUMMARY OF THE INVENTION

Accordingly it is an object of the invention to provide method and apparatus for speech coding using nonacoustic information in combination with acoustic information.

It is also an object of the invention to provide method and apparatus for speech coding using Electromagnetic (EM) wave generation and detection modules in combination with acoustic information.

It is also an object of the invention to provide method and apparatus for speech coding using radar in combination with acoustic information.

It is another object of the invention to use micropower impulse radar in conjunction with acoustic information for speech coding.

It is another object of the invention to use the methods and apparatus provided for speech coding for the purposes of speech recognition, mathematical approximation, information storage, speech compression, speech synthesis, vocoding, speaker identification, prosthesis, language teaching, speech correction, language identification, and other speech related applications.

The invention is a method and apparatus for joining nonacoustic and acoustic data. Nonacoustic information describing speech organs is obtained using Electromagnetic (EM) waves such as RF waves, microwaves, millimeter waves, infrared or optical waves at wavelengths that reach the speech organs for measurement. Their information is combined with conventional acoustic information measured with a microphone. They are combined, using a deconvolving algorithm, to produce more accurate speech coding than

obtainable using only acoustic information. The coded information, representing the speech, is then available for speech technology applications such as speech compression, speech recognition, speaker recognition, speech synthesis, and speech telephony (i.e., vocoding).

Simultaneously obtained EM sensor and acoustic information are used to define a time frame and to obtain the details of a human speaker's excitation function and vocal tract function for each speech time frame. The methods make available the formation of numerical feature vectors for characterizing the acoustic speech unit spoken each speech time frame. This makes possible a new method of speech characterization (i.e., coding) using a more complete and accurate set of information than has been available to previous workers. Such coding can be used for purposes of more accurate and more economical speech recognition, speech compression, speech synthesis, vocoding, speaker identification, teaching, prosthesis, and other applications.

The present invention enables the user to obtain the transfer function of the human speech system for each speech time frame defined using the methods herein. In addition, the present invention includes several algorithmic methods of coding (i.e., numerically describing) these functions for valuable applications in speech recognition, speech synthesis, speaker identification, speech transmission, and many other applications. The coding system, described herein, can make use of much of the apparatus and data collection techniques described in the copending patent application Ser. No. 08/597,596, filed 2/6/96, including EM wave generation, transmission, and detection, as well as data averaging, and data storage algorithms. The procedures defined in the copending patent application are called NASR or NonAcoustic Speech Recognition. Procedures based upon acoustic prior art are called CASR for Conventional Acoustic Speech Recognition, and these procedures are also used herein to provide processed acoustic information.

The following terms are used herein. An acoustic **speech unit** is the single or multiple sound utterance that is being described, recognized, or synthesized using the methods herein. Examples include syllables, demi-syllables, phonemes, phone-like speech units (i.e., PLUs),

diphones, triphones, and more complex sound sequences such as words. Phoneme acoustic-speech-units are used for most of the speech unit examples herein. A **speech frame** is a time during which speech organ conditions (including repetitive motions of the vocal folds) and the acoustic output remain constant within pre-defined values that define the constancy. Multiple time frames are a sequence of time frames joined together in order to describe changes in acoustic or speech organ conditions as time progresses. A **speech period**, or **pitch period** is the time the glottis is open and the time it is closed until the next glottal cycle begins, which include transitions to unvoiced speech or to silence. A **speech segment** is a period of time of sounded speech that is being processed using the methods herein. Glottal tissue includes vocal fold tissue and surrounding tissue, and glottal open/close cycles are the same as vocal fold open/close cycles. The word **functional**, as used herein, means a mathematical function with both variables and symbolic parameter-coefficients, whereas the word **function** means a functional with defined numerical parameter-coefficients.

The present methods and apparatus work for all human speech sounds and languages, as well as for animal sounds generated by vocal organ motions detectable by EM sensors and processed as described. The examples are based on, but not limited to American English speech.

1) EM Sensor Generator:

All configurations of EM wave generation and detection modules that meet the requirements for frequency, timing, pulse format, tissue transmission, and power (and safety) can be used. EM wave generators may be used which, when related to the distance from the antenna(s), operate in the EM near-field mode (mostly non-radiating), in the intermediate-EM-field mode where the EM wave is both non-radiating and radiating, and in the radiating far-field mode (i.e. most radars). EM waves in several wavelength-bands between $<10^8$ to $>10^{14}$ Hz can penetrate tissue and be used as described herein. A particular example is a wide-band microwave EM generator impulse radar, radiating 2.5 GHz signals and repeating its measurement at a 2 MHz pulse repetition rate, which penetrates over 10 cm into the head or

neck. Such units have been used with appropriate algorithms to validate the methods. These units have been shown to be economical and safe for routine human use. The speech coding experiments have been conducted using EM wave transmit/receive units (i.e., impulse radars) in two different configurations. In one configuration, glottal open-close information, together with simultaneous acoustic speech information, was obtained using one microphone and one radar unit. In a second set of experiments, three EM sensor units and one acoustic unit were used. In addition, a particular method is described for improving the accuracy of transmitting and receiving an electromagnetic wave into the head and neck, for very high accuracy excitation function descriptions.

2) EM Sensor Detector:

Many different EM wave detector modes have been demonstrated for the purpose of obtaining nonacoustic speech organ information. A multiple pulse, fixed-range-gate reception system (i.e., field disturbance mode) has been used for vocal fold motion and nearby tissue motion detection. Other techniques have been used to determine the positions of other vocal organs to obtain added information on the condition of the vocal tract. Many other systems are described in the radar literature on EM wave detection, and can be employed.

3) Configuration structures and Control System:

Many different control techniques for portable and fixed EM sensor/acoustic systems can be used for the purposes of speech coding. However, the processing procedures described herein may require additional and different configurations and control systems. For example, in applications such as high fidelity, "personalized" speech synthesis, extra emphasis must be placed on the quality of the instrumentation, the data collection, and the sound unit parsing. The recording environments, the instrumentation linearity, the dynamic range, the relative timing of the sensors (e.g. acoustic propagation time from the glottis to the microphone), the A/D converter accuracy, the processing algorithms' speed and accuracy, and the qualities of play back instrumentation are all very important.

4) Processing Units and Algorithms :

For each set of received EM signals and acoustic signals there is a need to process and extract the information on organ positions (or motions) and to use the coded speech sounds for the purposes of deconvolving the excitation from the acoustic output, and for tract configuration identification. For example, information on the positions of the vocal folds (and therefore the open area for air flow) vs. time is obtained by measuring the reflected EM waves as a function of time. Similarly, information on the conditions of the lips, jaw, teeth, tongue, and vellum positions can be obtained by transmitting EM waves from other directions and using other pulse formats. The reflected and received signals from the speech organs are stored in a memory and processed every speech time frame, as defined below. The reflected EM signals can be digitized, averaged, and normalized, as a function of time, and feature vectors can be formed.

The present invention uses EM sensor data to automatically define a speech time frame using the number of times that the glottis opens and closes for vocalized speech, while the conditions of other speech organs and the acoustics remain substantially constant. The actual speech time frame interval used for the processing (for either coding or reconstructing) can be adapted to optimize the data processing. The interval can be described by one or several constant single pitch periods, by a single pitch period value and a multiplier describing the number of substantially identical periods over which little sound change occurs, or it can use the pitch periods to describe a time interval of essentially constant speech but with "slowly changing" organ or acoustic conditions. The basic glottal-period timing-unit serves as a master timing clock. The use of glottal periods for master timing makes possible an automated speech and vocal organ information processing system for coding spoken speech, for speech compression, for speaker identification, for obtaining training data, for codebook or library generation, for synchronization with other instruments, and for other applications. This method of speech frame definition is especially useful for defining diphones and higher order multiple sound acoustic speech units, for time compression and alignment, for speaker speech rate normalization, and for prosody parameter definition and

implementation. Timing can also be defined for unvoiced speech, similarly to the procedures used for vocalized speech.

Once a speech time frame is defined, the user deconvolves the acoustic excitation function from the acoustic output function. Both are simultaneously measured over the defined time frame. Because the mathematical problems of "invertability" are overcome, much more accurate and efficient coding occurs compared to previous methods. By measuring the human excitation source function in real time, including the time during which the vocal folds are closed and the airflow stops (i.e., the glottal "zeros"), accurate approximations of these very important functional shapes can be employed to model each speech unit. As a result of this new capability to measure the excitation function, the user can employ very accurate, efficient digital signal processing techniques to deconvolve the excitation function from the acoustic speech output function. For the first time, the user is able to accurately and completely describe the human vocal tract transfer function for each speech unit.

There are three speech functions that describe human speech: **E(t)** = excitation function, **H(t)** = transfer function, and **I(t)** = output acoustics function. The user can determine any one of these three speech functions by knowing the two other functions. The human vocal system operates by generating an excitation function, **E(t)**, which produces rapidly pulsating air flow (or air pressure pulses) vs. time. These (acoustic) pulses are convolved with (or filtered by) the vocal tract transfer function, **H(t)**, to obtain a sound output, **I(t)**. Being able to measure, conveniently in real time, the input excitation E and the output I, makes it possible to use linear mathematical processing techniques to deconvolve E from I. This procedure allows the user to obtain an accurate numerical description of the speaker's transfer function H. This method conveniently leads to a numerical Fourier transform of the function H, which is represented as a complex amplitude vs. frequency. A time domain function is also obtainable. These numerical functions for H can be associated with model functions, or can be stored in tabular form, in several ways. The function H is especially useful because it describes, in detail, each

speaker's vocal tract acoustical system and it plays a dominant role in defining the individualized speech sounds being spoken.

Secondly, a synthesized output acoustic function, **I(t)**, can be produced by convolving the voiced excitation function, **E(t)**, with the transfer function, **H(t)**, for each desired acoustic speech unit. Thirdly, the excitation function, E, can be determined by deconvolving a previously obtained transfer function, **H**, from a measured acoustic output function, **I**. This third method is useful to obtain the modified-white-noise excitation-source spectra to define an excitation function for each type of unvoiced excitation. In addition, these methods can make use of partial knowledge of the functional forms E, **H**, or **I** for purposes of increasing the accuracy or speed of operation of the processing steps. For example, the transfer function **H** is known to contain a term **R** which describes the lips-to-listener free space acoustic radiation transfer function. This function R can be removed from **H** leaving a simpler function, H^*, which is easier to normalize. Similar knowledge, based on known acoustic physics, and known physiological and mechanical properties of the vocal organs, can be used to constrain or assist in the coding and in specific applications.

<u>The Bases of the Methods</u>:

1) The vocalized excitation function of a speaker and the acoustic output from the speaker are accurately and simultaneously measured using an EM sensor and a microphone. As one important consequence, the natural opening and closing of a speaker's glottis can serve as a master timing clock for the definition of speech time frames.

2) The data from 1) is used to deconvolve the excitation function from the acoustic output and to obtain the speaker's vocal tract transfer function each speech time frame.

3) Once the excitation function, the transfer function, and the acoustic function parameters are determined, the user forms feature vectors that characterize the speech in each time frame of interest to the degree desired.

4) The formation procedures for the feature vectors are valuable and make possible new procedures for more accurate, efficient, and economical speech coding, speech compression, speech recognition,

speech synthesis, telephony, speaker identification, and other related applications.

Models and Coding of Human Speech:

It is common practice in acoustic speech technology as well as in many linear system applications to use mathematical models of the system. Such models are used because it is inefficient to retain all of the information measured in a time-evolving (e.g., acoustic) signal, and because they provide a defining constraint (e.g., a pattern or functional form) for simplifying or imposing physical knowledge on the measured data. Users want to employ methods to retain just enough information to meet the needs of their application and to be compatible with the limitations of their processing electronics and software. Models fall into two general categories--linear and non-linear. The methods herein describe a large number of linear models to process both the EM sensor and the acoustic information for purposes of speech coding that have not been available to previous practitioners of speech technology. The methods also include coding using nonlinear models of speech that are quantifiable by table lookup or by curve fitting, by perturbation methods, or using more sophisticated techniques relating an output to an input signal, that also have not been available to users.

The simultaneously obtained acoustic information can also be processed using well known standard acoustic processing techniques. Procedures for forming feature vectors using the processed acoustic information are well known. The resulting feature vector coefficients can be joined with feature vectors coefficients generated by the EM sensor/acoustic methods described herein.

Vocal system models are generally described by an excitation source which drives an acoustic resonator tract, from whence the sound pressure wave radiates to a listener or to a microphone. There are two major types of speech: 1) voiced where the vocal folds open and close rapidly, at approximately 70 to 200 Hz, providing periodic bursts of air into the vocal tract, and 2) "unvoiced" excitations where constrictions in the vocal tract cause air turbulence and associated modified-white acoustic-noise. (A few sounds are made by both processes at the same time).

The human vocal tract is a complex acoustic-mechanical filter that transforms the excitation (i.e., noise source or air pressure pulses) into recognizable sounds, through mostly linear processes. Physically the human acoustic tract is a series of tubes of different lengths, different area shapes, with side branch resonator structures, nasal passage connections, and both mid and end point constrictions. As the excitation pressure wave proceeds from the excitation source to the mouth (and/or nose), it is constantly being transmitted and reflected by changes in the tract structure, and the output wave that reaches the lips (and nose) is strongly modified by the filtering processes. In addition, the pressure pulses cause the surrounding tissue to vibrate at low levels which affects the sound as well. It is also known that a backward propagating wave (i.e. reflecting wave off of vocal tract transitions) does travel backward toward the vocal folds and the lungs. It is not heard acoustically, but it can influence the glottal system and it does cause vocal tract tissue to vibrate. Such vibrations can be measured by an EM sensor used in a microphone mode.

Researchers at Bell Laboratories (Flanagan, Olive, Sondhi and Schroeter ibid.) and elsewhere have shown that accurate knowledge of the excitation source characteristics and the associated vocal tract configurations can uniquely characterize a given acoustic speech unit such as a syllable, phoneme, or more complex unit. This knowledge can be conveyed by a relatively small set of numbers, which serve as the coefficients of feature vectors that describe the speech unit over each speech time frame. They can be generated to meet the degree of accuracy demanded by the applications. It is also known that if a change in a speech sound occurs, the speaker has moved one or more speech organs to produce the changed sound. The methods described herein can be used to detect such changes, to define a new speech time frame, and to form a new feature vector to describe the new speech conditions.

The methods for obtaining accurate vocal tract transfer function information can be used to define coefficients that can be used in the feature vector that describes the totality of speech tract information for each time frame.

One type of linear model often used to describe the vocal tract transfer function is an acoustic-tube model (see Sondhi and Schroeter, ibid). A user divides up the human vocal tract into a large number of tract segments (e.g., 20) and then, using advanced numerical techniques, the user propagates (numerically) sound waves from an excitation source to the last tract segment (i.e., the output) and obtains an output sound. The computer keeps track of all the reflections, re-reflections, transmissions, resonances, and other propagation features. Experts find the sound to be acceptable, once all of the parameters defining all the segments plus all the excitation parameters are obtained.

While this acoustic tube model has been known for many years, the parameters describing it have been difficult to measure, and essentially impossible to obtain in real time from a given speaker. The methods herein, describing the measuring of the excitation function, the acoustic output, and the deconvolving procedures yields a sufficient number of the parameters needed that the constrictions and conditions of the physical vocal tract structure model can be described each time. One-dimensional numerical procedures, based upon time-series techniques, have been experimentally demonstrated on systems with up to 20 tract segments to produce accurate models for coding and synthesis.

A second type of linear acoustic model for the vocal tract is based upon electrical circuit analogies where excitation sources and transfer functions (with poles and zeros) are commonly used. The corresponding circuit values can be obtained using measured excitation function, output function, and derived transfer-function values. Such circuit analog models range from single mesh circuit analogies, to 20 (or more) mesh circuit models. By defining the model with current representing volume-air-flow (and voltage representing air pressure), then using capacitors to represent acoustic tract-section chamber-volumes, inductors to represent acoustic tract-section air-masses, and resistors to represent acoustic tract-section air-friction and heat loss values, the user is able to model a vocal tract using electrical system techniques. Circuit structures (such as T's and/or Pi's) correspond to the separate structures of the acoustic system, such as tube lengths, tongue positions, and side resonators of a particular individual. In principle,

the user chooses the circuit constants and structures to meet the complexity requirements and forms a functional, with unknown parameter values. In practice it has been easy to define circuit analogs, but very difficult to obtain the values describing a given individual and even more difficult to measure them in real time. Using a one mesh model, an electrical analog method has been experimentally validated for obtaining the information needed to determine the feature vector coefficients of a human in real time.

A third important model is based upon time series procedures (a type of digital signal processing) using autoregressive, moving average (ARMA) techniques. This approach is especially valuable because it characterizes the behavior of a wave as it traverses a series of transitions in the propagating media. The degree of the ARMA functional reflects the number of transitions (i.e., constrictions and other changes) in acoustic tracts used in the model of the individual. Such a model is also very valuable because it allows the incorporation of several types of excitation sources, the reaction of the propagating waves on the vocal tract tissue media itself, and the feedback by backward propagating wave to the excitation functions. The use of ARMA models has been validated using 14 zeros and 10 poles to form the feature vector for the vocal tract transfer function of a speaker saying the phoneme /ah/ as well as other sounds.

A fourth method is to use generalized curve fitting procedures to fit data in tables of the measured excitation-function and acoustic-output processed values. The process of curve fitting (e.g., using polynomials, LPC procedures, or other numerical approximations) is to use functional forms that are computationally well known and that use a limited number of parameters to produce an acceptable fit to the processed numerical data. Sometimes the functional forms include partial physical knowledge. These procedures can be used to measure and quantify arbitrary linear as well as non-linear properties relating the output to the input.

<u>5) Speech Coding System and Post Processing Units:</u>

The following devices can be used as part of a speech coding system or all together for a variety of user chosen speech related

applications. All of the following devices, except generic peripherals, are specifically designed to make use of the present methods and will not operate at full capability without these methods.

a) Telephone receiver/transmitter unit with EM sensors: A unit, chosen for the application, contains the needed EM sensors, microphone, speaker, and controls for the application at hand. The internal components of such a telephone-like unit can include one or more EM sensors, a processing unit, a control unit, a synthesis unit, and a wireless transmission unit. -This unit can be connected to a more complex system using wireless or transmission line techniques.

b) Control Unit: A specific device that carries out the control intentions of the user by directing the specific processors to work in a defined way, it directs the information to the specified processors, it stores the processed data as directed in short or long term memory, it can transmit the data to another specified device for special processing, to display units, or to a communications devices as directed.

c) Speech Coding Unit: A specific type of a coding processor joins information from an acoustic sensor to vocal organ information from the EM sensor system (e.g., from vocal fold motions) to generate a series of coefficients that are formed into a feature vector for each speech time frame. The algorithms to accomplish these actions are contained therein.

d) Speech Recognizer: Post processing units are used to identify the feature vectors formed by the speech coding unit for speech recognition applications. The speech recognition unit matches the feature vector from c) with those in a pre-constructed library. The other post-processing units associated with recognition (e.g., spell checkers, grammar checkers, and syntax checkers) are commonly needed for the speech coding applications.

e) Speech Synthesizer and Speaker: Coded speech can be synthesized into audio acoustic output. Information, thus coded, can be retrieved from the user's recent speech, from symbolic information (e.g., ASCII symbol codes) that is converted into acoustic output, from information transmitted from other systems, and from system

communications with users. Furthermore, the coded speech can be altered and synthesized into many voices or languages.

f) Speaker Identification: As part of the post processing, the idiosyncratic speech and organ motion characteristics of each speaker can be analyzed and compared in real time. The comparison is to known records of the speaker's physical speech organ motions, shapes, and language usage properties for a sequence of words. The EM sensor information adds a new dimension of sophistication in the identification process that is not possible using acoustic speech alone.

g) Encryption Units: Speech coded by the procedures herein can be further coded (i.e., encrypted) in various ways to make them difficult to use by other than an authorized user. The methods described herein allow the user to code speech, with such a low bandwidth requirement, that encryption information can be added to the transmitted speech signal without requiring additional bandwidth beyond what is normally used.

h) Display Units: Computer rendered speech information must be made available to the user for a variety of applications. A video terminal is used to show the written word rendition of the spoken words, graphical renditions of the information, (e.g., the articulators in a vocal tract), a speaker is used to play previously recorded and coded speech to the user. The information can be displayed by printed using printers or fax machines.

i) Hand Control Units: Hand control units can assist in the instruction of the system being spoken to. The advantage of a hand control unit (similar to a "mouse") is that it can assist in communicating or correcting the type of speech being inputted. Examples are to distinguish control instructions from data inputting, to assist in editing by directing a combined speech-hand-directed cursor to increase the speed of identifying displayed text segments, to increase the certainty of control by the user, to elicit play-back of desired synthesized phrases, to request vocal tract pictures of the speakers articulator positions for language correction, etc.

j) Language Recognizer and Translator Unit: As the speaker begins to talk into a microphone, this device codes the speech

and characterizes the measured series of phonemes as to the language to which they belong. The system can request the user to pronounce known words which are identified, or the system can use statistics of frequent word sound patterns to conduct a statistical search through the codebooks for each language.

It is also convenient to use this same unit, and the procedures described herein, to accept speech recognized words from one language and to translate the symbols for the same words into the speech synthesis codes for the second language. The user may implement control commands requesting the speaker to identify the languages to be used. Alternatively, the automatic language identification unit, can use the statistics of the language, to identify the languages from which and to which the translations are to take place. The translator then performs the translation to the second desired language, by using the speech unit codes, and associated speech unit symbols, that the system generates while the first language is spoken. The speech codes, generated by the translator, are then converted into symbols or into synthesized speech in the desired second language.

k) Peripheral Units: Many peripheral units can be attached to the system as needed by the user making possible new capabilities. As an example, an auxiliary instrument interface unit allows the connection of instruments, such as a video camera, that require synchronization with the acoustic speech and speech coding. A communications link is very useful because it provides wireless or transmission line interfacing and communication with other systems. A keyboard is used to interface with the system in a conventional way, but also to direct speech technology procedures. Storage units such as disks, tape drives, semiconductor memories are used to hold processed results or, during processing, for temporary storage of information needed.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a schematic diagram of a prior art open loop acoustic speech coding system.

Fig. 2 is a schematic diagram of a combined nonacoustic/acoustic speech coding system using an EM sensor and a microphone, including optional auxiliary instruments.

Fig. 3A shows a schematic diagram of a highly accurate and flexible vocal tract laboratory measuring system for speech coding.

Fig. 3B shows a system for speech coding using three micropower radars and an acoustic microphone.

Fig. 4 shows an EM sensor directing EM radiation into the neck of a speaker with vocal folds shown in an open condition.

Fig. 5 is a flow chart showing the processing of simultaneously recorded acoustic data and EM sensor data, and subsequent deconvolution.

Fig. 6 is an acoustic and air flow model of vocal system showing an EM sensor for vocal folds and a microphone acoustic detector.

Fig. 7 is a continuous model of the vocal tract divided into 20 segments.

Fig. 8 is a schematic diagram of a speech coding system using EM sensors and acoustic data.

Figs. 9A,B are time domain data for the speech sound /ah/ using an acoustic pressure sensor and an EM glottal tissue sensor.

Figs. 10A,B are Fourier power spectra for the acoustic microphone data and the EM sensor measurements of glottal cycles for the sound /ah/.

Fig. 11A shows Fourier transfer function amplitude coefficients obtained for the two-tube phoneme /ah/.

Fig. 11B shows Fourier transfer function amplitude coefficients obtained for the single tube phoneme /ae/.

Fig. 12A shows a feature vector for the phoneme /ah/,

Fig. 12B shows the ARMA poles and zeros for Fig. 9A.

Fig. 12C shows the corresponding ARMA "a"'s and "b"'s for the sound /ah/ represented in Fig. 11A.

Figs. 13A-F show images of vocal folds opening and closing during one speech frame period, and characteristic dimensions.

Figs. 14A,B show the substantially simultaneously recorded acoustic signal and the corresponding EM sensor signal showing glottal motion versus time for the phoneme /ah/.

Fig. 15A shows several acoustic speech segments for the word "lazy".

Fig. 15B shows speech time frames and EM sensor vocal fold signals for the voiced and combination voiced/unvoiced unit /z/ in the word "lazy".

Fig. 16 is a source and impedance model that is an electrical analog to an acoustic model.

Fig. 17A shows a single mesh electrical analog circuit that models the first formant of the sound /ae/, using volume air flow as the independent variable.

Fig. 17B shows a single mesh electrical analog circuit that uses air pressure as the independent variable.

Fig. 18A shows a method of normalizing a speaker dependent feature vector coefficient, ${}_{meas}c_n$, to a normalized coefficient, ${}_{normal}c_n$.

Fig. 18B shows a method of quantization of a normalized coefficient into one quantized value that represents a quantized band of coefficients, over which no important sound changes occur.

Fig. 19 shows the comparison between the measured and synthesized power spectra of the acoustic speech phoneme /ah/.

Fig. 20 shows a telephone hand-set vocoding apparatus with receiver-speaker and microphone, including EM sensors for coding, and a synthesizer for decoding.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

General Principles

Figure 2 shows a speech processing model based on an EM sensor that is used to measure the motions of vocal fold interfaces and glottal tissue. These motions can be related to the volume air flow or glottal pressure, and can be measured simultaneously with the accompanying speech. Knowledge of the voiced excitation input and the acoustic output of a human vocal tract provides sufficient information to accurately deconvolve the excitation from the output.

The information from the sensors and from the deconvolving process makes possible new methods to code human speech in real time, and in an economical, safe, convenient, and accurate manner.

In Figure 2, signals from an acoustic microphone 1 are processed in block 2 where the acoustic signals are digitized and feature vectors are formed for selected time frames. Electromagnetic signals from EM vocal fold sensor 3 are input into processing block 4 where the signals are digitized and time units are defined and feature vectors are formed. The acoustic and EM feature vectors from processing blocks 2 and 4 are input into processing block 5 where the EM signal is deconvolved from the acoustic signal. Processing unit 4 also controls timing unit 6, which sets the master timing and speech time frames, and which is connected back to processing units 2 and 4. The deconvolved output from unit 5 is input into unit 7 where the data is fit to a transfer function, which is used to form a joint feature vector in unit 8, which is then stored in a memory or code book in block 9. Optionally, additional EM sensors 10 can be used to measure vocal tract conditions and other sensors 11 can also be utilized. Feature vectors from sensors 10, 11 are formed in blocks 12, 13 and the best transfer function for deconvollution is selected in block 14, which is then input into unit 7. In addition, feature vectors from block 2 can be sent directly to a CASR (conventional acoustic recognition system), and feature vectors from blocks 12,13 can be sent via block 15 for separate processing and subsequent use in the applications described herein.

Figures 3A and Figure 3B show two types of laboratory apparatus for measuring the simultaneous properties of several speech organs using EM sensors and for obtaining simultaneous acoustic information. Figure 3A, in particular, shows highly accurate laboratory instrumentation assembled to obtain very high fidelity, linear, and very large dynamic range information on the vocal system during each speech time frame. Figure 3A shows a view of a head with three antennas 21, 22, 23 and an acoustic microphone 24 mounted on a support stand 25. Antennas 21, 22, 23 are connected to pulse generators 26a, b, c through transmit/receiver switches 27a, b, c respectively. Pulse generators 26a, b, c apply pulses to antennas 21, 22, 23, which are directed

to various parts of the vocal system. Antennas 21, 22, 23 pick up reflected pulses, which are then transmitted back through switches 27a, b, c to pulse receivers and digitizers (e.g., sample and hold units) 28a, b, c. Acoustic information from microphone 24 is also input into pulse receiver and digitizer 28d. Support stand 25 positions the antennas 21, 22, 23 to detect signals from various parts of the vocal tract, e.g., by using face positioning structure 29 and chest positioning structure 30. As shown, antenna 21 is positioned to detect the tongue, lip, velum, etc. Antenna 22 is positioned to detect tongue and jaw motion and antenna 23 is position to detect vocal fold motion.

Figure 3B shows how presently available micro-impulse radars have been used to obtain valuable speech organ information in a controlled setting. The EM sensor signals from these EM sensors, measuring vocal fold or other tissue motion, are related to the true voiced excitation signal (i.e. volume air flow vs. time or pressure versus time) using the methods herein. Figure 3B shows a view of a head with three EM sensor transmit/receive modules 31, 32, 33 and an acoustic microphone 34 mounted on a support stand 35. The configuration is similar to that in Figure 3A except that entire EM motion sensors 31, 32, 33 are mounted on the stand 35 instead of just antennas with the remaining associated electronics being mounted in a remote rack. Many experiments referenced in this patent application were conducting using apparatus similar to that shown in Fig. 3B.

Figure 4 shows how an EM wave from an electromagnetic wave generator is used to measure the conditions of the vocal folds in a human speaker's neck. The wave is shown as radiated from the antenna; however other measuring arrangements can use an EM wave in the near field or in the intermediate field, in addition to the far field radiated EM wave as used in most radars. The EM wave is generated to measure the conditions of the vocal folds and the glottal tissue surrounding the vocal fold structure as often and as accurately as needed for the accuracy of the application.

Figure 5 shows a system in which knowledge of the vocalized excitation function is used to deconvolve the speech vocal tract transfer function information from measured acoustic speech

output each time frame. All of the information gathered during each speech time frame, including acoustics, EM sensor information, and deconvolved transfer function information, can be processed, normalized, quantized, and stored (along with control information) in a feature vector representing the speaker's voice during one or more speech time frames. Similar deconvolving procedures are used with unvoiced excitation functions. As shown in Figure 5, an EM sensor control unit 40 drives a repetition rate trigger 41, which drives pulse generator 42, which transmits one or more pulses from antenna 43. EM sensor control unit 40 sets the pulse format, time frame interval, integration times, memory locations, function forms, and controls and initializes pulse generator 42. Control unit 40 and trigger 41 also actuate switch 45 through delay 44 to range gate received pulses. Antenna 43 is positioned to direct transmitted pulses towards the vocal organs and receive pulses reflected therefrom. The received pulses pass through switch 45 and are integrated by integrator 46, then amplified by amplifier 47, and passed through a high pass filter 48 to a processing unit 49. Processing unit 49 contains an AD converter for digitizing the EM signals and also includes zero location detector, memory detector, and obtains glottal area versus time. The digitized and processed data from unit 49 is stored in memory bins 50, from which excitation function feature vectors are formed in block 51. Simultaneously, signals from an acoustic microphone 52 are digitized by AD converter 53, which is also controlled and synchronized by EM sensor control unit 40. The digitized data from AD converter 53 is stored in memory bins 54 from which acoustic feature vectors are formed in block 55. The digitized vocal fold data from memory bins 50 is used to produce a glottal Fourier transform 56, while the digitized acoustic data in memory bin 54 is used to produce an acoustic Fourier transform 57. The two Fourier transforms 56, 57 are deconvolved in block 58 to produce a vocal tract Fourier transform 59 which is then fit to a prechosen functional form to form a vocal tract feature vector in block 60.

Figure 6 shows a schematic of the human vocal system from an acoustic perspective. Figure 6 also identifies the major components utilized in speech, with an EM sensor 61 positioned to

detect glottal motions (including those of the vocal folds) which form an excitation source for the vocal tract, and an acoustic sensor 62 positioned to receive acoustic output from the mouth. The physical behavior of acoustic excitation pulses, after they are generated by the vocal folds or after generation at air passage constrictions, and as they traverse and are filtered by the varying tubes and chambers, are measured as acoustic pressure waves by the acoustic sensor (e.g., a microphone). Procedures described herein show how to describe the consequences of all of the important vocal tract structures, how to determine when they change to form a new sound, and how to code such condition for subsequent applications. The condition of the human speech organ structure is known to provide sufficient information to identify the acoustic speech units being articulated by that structure. In addition, it is known that these structures vary from individual to individual, and the way they are shaped and moved to articulate a sequential series of acoustic speech units varies from language to language and from individual to individual. Knowledge of such individual structural patterns, and their time sequencing to form speech sounds, forms the basis for speaker identification and language identification.

Figure 7 is a sketch of a cut through a human vocal system showing transverse dimensions along the center plane. The dotted lines and numbers show where one might approximate the vocal tract by short approximately circular cylinder constant sections. At each dotted interface, the cylinder would change diameter and, thus, a propagating acoustic wave from the glottis to the lips and/or nose would be both transmitted and reflected. In human vocal systems a cross section is not circular and the transitions are smooth. By segmenting this structure into a sufficient number of sub-structures (e.g., 20), each having a small dimensional change from the neighbors, accurate descriptions of the air flow (and pressure) can be obtained. Well known numerical and/or time series (e.g., ARMA) techniques have been used to describe the acoustic wave as it propagates from the excitation source to the microphone (or human ear) detector. Time series analysis (e.g. Z transform) procedures are especially useful for characterizing such

systems, because their functional forms easily accommodate a series of reflecting and transmitting structures. They are used herein to describe many of the transfer function examples.

Figure 8 schematically illustrates a speech technology system 70 using sensor 71, which includes both EM sensors and acoustic detectors. Sensor 71 could be, for example, similar to the device shown in Figure 3B or built into a telephone receive/transmit unit as in Figure 20. Sensor 71 is connected by a wireless (RF or optical) link or cable communication line 72 to a coding unit 74, which has associated therewith a control unit 73. Coding unit 74 is connected to language recognizer and translator 75, speech synthesizer 76, speech recognizer 77, and word spelling/syntax/grammar generator 78. A hand control unit 79 is connected to coding unit 74. Control unit 73 is connected to coding unit 74 for switching units and for directing information flow. Other peripheral equipment can be connected to coding unit 74 through control unit 73. For example, a video terminal 80, a communications link 81 to wires, cellular, wireless, fiber optics, etc., an encryption unit 82, a speaker identification unit 83, an auxiliary instrument interface unit 84 with a video camera 85 connected thereto, a printer or fax 86, or a loud speaker 87 can all be connected to control unit 73. Such a system makes it possible to record and process speech information, to code the information, and to use this coded information for applications such as forming language codebooks, speech recognition, speech synthesis, speaker identification, vocoding, language identification, simultaneous translation, synchronization of speech with video systems and other instruments, low bandwidth coding and encryption, speech correction and prosthesis, and language learning.

The system represented in Fig. 8 can be simplified and miniaturized for special applications. For example, Fig. 20 shows a portable, specialized version for vocoding because it obtains EM sensor plus acoustic information, processes it, codes it, and sends it into a transmission system that carries the information to a similar handheld unit for decoding and synthesizing of speech for the listener.

Deconvolving the Vocal System Excitation Function:

This method has been demonstrated using the EM glottal opening (i.e., vocal fold) area information and acoustic information measured for one or several sequential speech time frame periods to deconvolve the vocal system volume air flow source function from the measured acoustic speech output from a human speaker. Figures 9A,B show raw acoustic microphone and glottal motion data. The Fourier transforms of the data can be obtained and are shown in Figures 10A,B. The numerical representations of these two functions allow the user to obtain a numerical representation (i.e., a complex number coefficient representation) of the transfer function representing the acoustic filtering of the human vocal tract during the time frame or frames. The deconvolving of the excitation function from the acoustic output can be accomplished using real time techniques, time series techniques, fast Fourier transform techniques, model based transform techniques, and other techniques well known to experts in the field of data processing and deconvolving. Examples are shown whereby the Fourier transform of the acoustic output is divided by the excitation function input. Figure 11A shows the two tube sound /ah/ derived by using inputs from Figures 9A,B and 10A,B. Figure 11B shows the transfer function for the single tube sound /ae/ which is deconvolved using acoustic and vocal fold data similar to that for the two tube sound /ah/.

By using other EM sensors (in addition to the glottal sensor) to determine other speech organ location information, with or without simultaneous acoustic data, one can determine the optimal transfer functional structure to use for best convergence or for most accurate fitting of the transfer function. (Herein, functional is used to mean a specific function form, but with unspecified constants). An example is to use a lip sensor to report that when the lips are closed, during the articulation of a nasal phoneme /m/, the transfer functional form must contain a spectral zero due to the closed mouth cavity.

An example is to choose an ARMA functional (i.e. time series) description, with an appropriate number of poles and zeros, for each speech time interval frame. The number of poles and zeros are chosen to represent the complexity of the model and the desired accuracy of the resultant coding.

I(t), and E(t) are the measured acoustic output and EM excitation respectively. The algebraic input/output relation using the transfer function H(z) in the z-transform variable is:

$$I(z) = H(z) * E(z)$$

where H(z) is given in factored, pole-zero form, by:

$$H(z) = \frac{(z-z_1)(z-z_2)(z-z_3)\cdots(z-z_m)}{(z-p_1)(z-p_2)(z-p_4)\cdots(z-p_n)}.$$

Equivalently, the transfer function, functional form, can be written in a/b notation, where a's and b's are the coefficients of the mth order numerator and nth order denominator polynomials respectively.

$$H(z) = \frac{b_o + b_1 z^{-1} + b_2 z^{-2} + b_3 z^{-3} + \cdots + b_m z^{-m}}{a_o + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + \cdots + a_n z^{-n}}.$$

By using well known deconvolving techniques for the ARMA functionals one can divide the transformed microphone acoustic pressure signal by the transformed excitation source signal (using complex numbers) and thereby obtain the amplitude and phase of the transfer function. The transfer function is defined by the poles and zeros, or by the a and b coefficients in the two different ARMA functionals shown above. Furthermore one can, if desired, deconvolve the well known lip to microphone radiation function from the microphone signal to obtain the volume air flow function or transfer function at the lip and nose orifices. The ARMA approach, together with appropriate functional definitions of the excitation function and the acoustic data, makes possible the straightforward and automatic definition of a speech feature vector each speech time segment. For example, the algorithm stores the excitation function parameters defining a triangular approximation of the glottal volume air-flow versus time, it stores the transfer function using 14 poles and 10 zeros, the time frame duration, the prosody, some useful acoustic features, and the control values for subsequent speech technology purposes. For each of the functional forms, the information can be stored as a real time function, as a transformed function (e.g. Fourier transform) or as a mixed function as needed.

The feature vector information for each speech time frame can be normalized to a referenced speaker's (or speakers') feature vector

for the speech sound spoken in the time frame. The normalization method is to compare measured (and processed) vector coefficients to those from both the user and from the reference speaker. Those of the reference speaker have been recorded during earlier training sessions. Normalization also removes variations in the interaction between the EM-sensors and the individual qualities of each speaker, as well as variations from one unit of equipment to another. In addition, the continuous value-range of each individual's coefficients, which represent a vocal articulator's range, can be quanitized to a smaller number of values. The "quantized" values are chosen such that a change, from one quantized coefficient value to the next, represents a desired user-distinguishable effect on the application. An example is that each quantized coefficient value represents a just-discernible change in a synthetic speech sound. These methods, described below, make possible the formation of speaker independent featured vectors for each speech segment. The coefficients in each a vector can be time-length independent, pitch normalized, rate normalized, articulator amplitude normalized and quantized, and they contain important aspects of the acoustic information. The methods described herein, make possible great improvements in speech coding because of the completeness of the vocal system information, the accuracy of coding the speech, the speaker and instrument independence, and the computational simplicity of the associated algorithms.

<u>**Example of Time Frame Definition and Feature Vector Formation:**</u>

For a male speaker saying the sound unit /ah/ extending over a time segment of 300 ms, the speech acoustic sensor and the vocal fold signal from the EM sensor were sampled at 11 kHz. Figures 9A and 9B show real time acoustic and glottal amplitude versus time signals, respectively. A transfer function was computed every 10 ms with a 32 ms Hamming window. Complex spectra, using both acoustic and glottal motion channels, were obtained using a 256 point FFT (Fast Fourier Transform). An ARMA model was used to best fit the input and output data in a least mean squares sense. Fourteen poles and ten zeros achieved the best fit. Such ARMA coefficients contain both magnitude and phase information. Knowledge of the ARMA coefficients allowed

the construction of a feature vector describing the sound /ah/ for each 10 ms speech frame. Those essentially-identical speech frames were combined into a 300 ms multi-pitch-period speech time frame (i.e., thirty speech frames, each 10 ms were joined into one multi-time speech frame). The frequency response of the acoustic output and excitation input functions are shown in Fig. 10A,B respectively; and the computed transfer function amplitudes are shown in Fig. 11A. A similar process was used to generate the transfer function amplitudes for the sound /ae/, which are shown in Fig. 11B.

The feature vector shown in Fig. 12A for the sound /ah/, was constructed using a total of p feature vector coefficients, c_1 through c_p, to describe the processed data. In this example, c_1 is used to describe the type of transfer functions used, e.g. "1" means the use of an ARMA functional in the "pole" and "zero" formulation; c_2 describes the number of "poles" and c_3 describes the number of "zeros" used for the fitting; c_4 indicates the kind of speech unit being spoken, e.g. "0" means isolated phoneme; c_5 describes the type of connection to a preceding acoustic sound unit to be used, e.g. "0" means a connection to the silence phoneme is needed; c_6 describes the connection to the following unit, e.g. "0" means a connection to a following silence phoneme is needed; c_7 describes the 300 ms multi-frame speech segment envelope; c_8 is the pitch (e.g., 120 vocal fold cycles/sec.); and c_9 describes the bandwidth of the fundamental harmonic. Other feature vector coefficients that describe the relative ratios of the 2nd through the 10th harmonic power to the first harmonic, are taken from the power transform of the vocal excitation (Fig. 10B). In addition the fall of the harmonic excitation power per octave, above 1 kHz, can be described by a line with -12db/octave negative slope. The "pole" and "zero" coefficient data (Fig. 12B) are shown and stored as appropriate coefficients in the vector in Fig. 12A. The last coefficient c_p is the symbol for the sound, and the next to last c_{p-1} is acoustic information from a CASR or similar system which is the acoustic energy per frame. If the user desires to use the alternative formulation of the ARMA transfer functional, the "a" and "b" coefficients can be used (see Fig. 12C).

An alternative approach to describe the feature vector for the "long" speech segment /ah/ is to perform Fourier transformations each 8.3 ms (the period for 120 Hz excitation), and to join 36 individual pitch period frames into a 300 ms long multiple frame speech segment. A second alternative approach would be to take the Fourier transform of the entire 300 ms segment, since it was tested to be constant; however the FFT algorithm would need to handle the large amount of data. Because of the constancy of the acoustic phoneme unit /ah/, the user chose to define the 300 ms period of constancy first, and to then process (i.e., FFT) the repetitive excitation and output acoustic signal with a convenient 10 ms period 30 times, and then average the results.

As a test (see Section below on Speech Synthesis) a synthetic speech segment was reconstructed from information in a vector like the one shown in Fig. 12A. The vocal fold excitation function was first reconstructed using the harmonic amplitude and phase information to generate a source term over an interval of 100 ms. The excitation function was sampled at 11 kHz or higher. The time sampled sequence was used to drive the ARMA model specified by a difference equation with poles and zeros. The output of the ARMA model was used to reconstruct the speech sound /ah/ as shown in the section on Speech Synthesis (see Fig. 19), and a pleasing sound, /ah/, was generated and heard by the user.

Applications of Preferred Embodiment:

The procedures to define speech time segments and to form feature vectors allow many applications. First, the user-speaker or other speakers, who serve as references, are asked to speak into a sensing and recording system, such as are shown in Figs. 3A or 3B. Feature vectors are formed for all single unit sounds in a language (e.g. syllables, phonemes, PLUs, and acoustic speech units) and for as many multisound unit sounds (e.g., diphonemes, triphonemes, words, and phrases) as are needed by the user for the application. The identified feature vectors, for the speech segment, can be normalized and quantized as needed, and are stored in a codebook (i.e., library). The identification of the stored feature vectors can be done in several ways. They can be labeled by the frame position in a time sequence of frames

or be labeled by a master timing clock. They can be labeled using known labeling of each feature vector with user provided acoustic speech unit names (e.g. Fig. 12A, last coefficient, c_p = ah, describes the phoneme /ah/). They can also be automatically labeled using speech recognition to add the missing acoustic speech unit label to the feature vector for the speech segment. Because of the direct relationships between speech organ positions, their rates of motion, and the sound units produced, the methods described herein provide a more fundamental parametrization of vocal system conditions during speech than has been possible before. They make possible simplified but very accurate descriptions of single acoustic speech units, as well as descriptions of acoustic speech units that include multiple phonemes such as diphones, triphones, whole words, and other well known combinations.

Once the speech segments are identified and stored, many applications are possible. They include speech recognition, speech synthesis, vocoding for telephony, speech prosthesis and speech correction, foreign language identification and learning, and speaker identification. For speech recognition, the user can perform direct phonetic-template matching with previously stored feature vectors in a library for the purposes of automatic speech unit identification. Similarly, the user can use Hidden Markov Models, or neural networks, or joint or exclusive statistical techniques for the identification of one or several consecutively formed feature vectors using previously stored information. For purposes of speech reconstruction (i.e., speech synthesis) the coding procedures make possible the characterization of any individual speaker's sounds. Then, using methods for accurate synthesis of each speech segment, many speech segments are joined together. Synthesized speech can be altered as desired. Speaker identification and language identification are made possible because the speech coding reflects the specific properties of each user and the properties of the language the user is speaking.

<u>Voiced Excitation Function Description</u>

The preferred method is based upon air volume flow through the vocal tract as the independent variable and air pressure as the dependent variable. An EM sensor is positioned in front of the

throat at the location of the vocal box (i.e., larynx). It measures the change in EM wave reflection from the vocal folds and surrounding glottal tissue as they open and close. The user can determine the relative volume of air flow through the glottal opening during the voicing of each voiced acoustic speech unit. This allows one to measure and generate, in an automated fashion, an accurate voiced speech excitation function of any speaker and to define the speech time frame interval or intervals during which this function provides a constant, periodic repetitive excitation.

One demonstrated method is to measure the change in EM wave reflection level from the glottal region as the vocal folds open and close using a "field disturbance" EM sensor optimized for glottal tissue motion detection. By time filtering to allow a signal bandpass of approximately 50 Hz to >2 kHz, the voiced glottal signal is easily measured and separated from other signals in the neck and from those associated with slower body motions moving the sensor relative to the neck. The next step is to associate each reflection condition with the area opening of the glottis. The area measurement methods are based upon using known physics of EM wave scattering from dielectric materials, by using mechanical and physiological models of the glottal tissues, and by calibration of EM sensors signals against physical air flow and/or pressure sensors. Then a model of air flow vs area, based upon fluid dynamic principles, is used. For other applications, depending upon the coding fidelity of speech needed, the EM sensor can be optimized to generate more accurate data, wider bandwidth data, and data with increased linearity and dynamic range.

Generalized methods of obtaining the vocalized excitation function include procedures where the EM sensor amplitude versus time signal is calibrated against laryngoscope pictures of glottal area vs. time and/or air sensor amplitude vs. time signals (e.g., using air flow and/or air pressure sensors). One method uses a laryngoscope to optically photograph the area opening, versus time, simultaneously with the EM sensor measurement of the EM reflection signals. Figs. 13A-F are examples of vocal fold opening and closing images of the glottal area. Another method is to place air sensors in various vocal

tract locations to calibrate the EM sensor signals against absolute air flow versus time signals, or against pressure versus time signals. A direct functional relationship between an EM-sensor signal-amplitude at a given time and the associated air flow signal (or its dual pressure value) at the same time is obtained by measuring both substantially simultaneously under the needed conditions of use for the speech vocabulary in the application. These methods are especially valuable for obtaining the glottal open and closure times and the shape (i.e., derivatives) of the air flow versus time signal at the moments of glottal opening and closure for coding applications needed for speech synthesis applications. Normalization procedures are used to correct the signals, and the relationships are stored in a lookup table or codebook, or the relationships are approximated by model based or curve fitted functions. Thus for each EM-sensor signal value from glottal tissue, an airflow or air pressure value can be associated.

Experiments with excitation functions based upon air volume flow were conducted to validate the methods. The data are analytically described by using well known fluid flow equations, one of which was described by Flanagan 1965 ibid on p.41, equation 3.46. The resistance to airflow through the glottal opening, at constant lung pressure, is given in equation (1) below. The resistance **Rg** is equal to the difference in pressure on either side of the glottal opening (i.e. the transglottal pressure P_s) divided by the total air flow U (i.e. volume air flow). For this example, ρ= air density, l = length of glottal slit, and w = transverse opening of glottal slit (see Fig. 13B). The viscous term in Eq. (1) is neglected, because it is only needed for small openings, and was not used for the validation experiments.

(1) $R_g = P_s/U = (\text{viscous term}) + 0.875\ \rho U/2(lw)^2$

(2) $P_s = U * R_g$

(3) $P_s = 0.875\, \rho\, U^2 / 2(lw)^2$

(4) $U = (lw) * (P_s/0.438\, \rho)^{1/2}$

The change in the glottal opening area, lw, is proportional to the change in the EM wave reflection caused by the change in the local dielectric value as the glottal tissue material moves. This example uses the approximation that the reflected EM wave-signal changes in proportion

to the reduction in glottal tissue mass as the glottis opens. This interpretation works well for the "field disturbance" type of EM sensor used in the experimental examples. Using knowledge about the shape of the glottal opening, a further relationship is developed whereby the tissue mass of the opening is reduced in proportion to w, the glottal width, in equation (4). Thus measuring "w" directly with the field disturbance EM sensor (or by using other sensor systems such as a range gated EM sensor) the needed area value versus time is obtained. Then using equation (4), the needed volume air flow signal, U, versus time is obtained from the area value, lw. Figures 14 A,B show an experimentally obtained acoustic signal and the associated EM sensor signal from glottal tissue motions. Using the relationships just derived between the EM sensor signal and the volume air flow, U, and assuming constant transglottal pressure, P_s, the signal in Fig. 14B describes the relative volume air flow, U, versus time.

The simplified analytical approach, used above for modeling the air flow resulting from EM sensor measurements of the glottal tissue motions, is employed to demonstrate the effectiveness of having excitation function data, the clarity of the timing information, and the directness of the deconvolving process. The experiments assumed constant lung pressure and constant transglottal pressure during each speech frame in this description of a short speech segment. For most cases relative changes in air flow, U(t), are sufficient, and slowly changing lung pressure does not matter. However, if lung pressure is needed, an EM sensor can be employed to measure the lung volume change or diaphragm motion to determine relative lung volume change. In the cases of changing transglottal pressure over the needed measurement periods, methods are described below. In addition, the change in the amplitude envelope of acoustic speech generated over several glottal periods can be recorded in a feature vector, and provide a measure of relative change in air flow and thus in excitation amplitude. Such amplitude changes provide important prosodic information for speech recognition, speech synthesis, and are especially valuable for speaker identification procedures where

individualized intonation of identical spoken phrases is very idiosyncratic.

The procedures used volume air flow as the independent variable. However EM sensors optimized to sense the condition of other glottal tissues, as they respond to changes in volume air flow or to local pressure, can be used and their responses can be fed into an equation (i.e., algorithm) which will provide a volume or a pressure versus time vocalized speech source function for use in coding procedures.

<u>Air Flow Corrections Due to Post- and Trans-Glottal Pressure Variations:</u>

It is known that for most conditions, the glottal opening is a high impedance air flow orifice, meaning that the glottal impedance is substantially higher than the following post glottal impedance values. In this approximation, post-glottal vocal tract changes do not affect the transglottal pressure and the air flow through the glottal orifice. However, in more realistic approximations, such air flow changes can be important. The user may wish to describe, more accurately, the voiced excitation function, and may wish to use one of the following methods employing EM sensor signals plus noted algorithmic procedures. While the above model of the air flow through the glottal orifice assumed constant pressure on both sides of the vocal folds (i.e., constant transglottal pressure), the effects of a postglottal pressure change during the speech time frame can be estimated using well known approximation techniques from electrical analogies and from physical principles, or can be measured using tissue motions sensitive to local pressure. These pressure corrections can be important because, from Figure 16, when the post glottal pressure P_1 (represented as voltage V_1) becomes a significant fraction of the lung pressure P_o (represented as voltage V_o), then the use of glottal area to define volume air-flow function, U, breaks down. An improved expression with the necessary corrections must be used for applications where the highest quality excitation function characterization is needed, e.g. during "obstruent" articulation.

By using the EM-sensor for glottal motion, in a high sensitivity mode, the user can measure low amplitude vocal-fold tissue

motions (e.g., vibrations) that are known to be caused by air flow pressure changes. Such pressure fluctuations are caused, for example, by backward propagating acoustic signals. Vibrations that affect the glottal opening can be distinguished from other surrounding tissue vibrations being sensed by the same EM sensor. Fig. 14B shows examples of such vibrations which slightly modulate the peak envelope-amplitude signal of the glottal-opening versus time signal. These are known to be associated with acoustic pressure waves, because when the low frequency glottal envelope is electronically filtered away, leaving the higher frequency vibration signals, the latter can be amplified and sent to a loud speaker. The broadcasted signals are recognizable as being nearly identical to the acoustic speech recorded by the microphone. These signals are measured to be small, and calculations describing the magnitude of these effects also indicate them to be small in most cases. In applications where high coding fidelity is important and where the compliance of the glottal tissue is needed for mechanical models or for speaker identification, the following methods are used to provide the needed additional information. Seven methods are described for accommodating the variations in the glottal-air flow versus time, due to transglottal pressure changes. They are used to form improved vocalized excitation function descriptions over the defined time frames of interest:

1) Make no changes to the glottal opening signal, even though it is known that the air flow model is being perturbed by changes in the transglottal pressure. Form a numerical approximation of the volume air flow function vs. time assuming constant transglottal pressure. Deconvolve the volume air flow function from the acoustic signal. Using an appropriate transform functional, find the numerical coefficients describing the transform function for the time frame. Construct a feature vector for the time frame, using the uncorrected excitation function, the related transfer function, and measured acoustic signal parameters (as well as other coefficients described below under feature vector formation). The three speech functions used in this method, E(t), H(t), and I(t) are together self-consistent. They can be used for real time feature vector formation and time frame definition, as well

as to generate the needed application specific codebooks realizing that many of the feature vector parameters (and thus the codebooks) are imperfect but they are all self-consistent. For many applications, feature vectors generated using this method are good enough.

2) Using physiological data of the individual speaker (or using an average human vocal tract) together with an air flow speech model of the transfer function, calculate the post glottal pressure from the impedance of the transfer function looking from the glottis forward. This procedure is well known to experts who model air flow and pressure in speech tracts. (An additional EM sensor to measure various vocal tract organ positions can be used to provide data to aid in choosing a transfer functional and its consequent impedance). Use this impedance to make a first order correction to the transglottal air pressure and thus a correction to the air flow obtained from Equations 1-4 above. Use the corrected volume air flow to form a corrected excitation function feature vector.

3) Remove post-glottal pressure induced vibrations of glottal tissue and nearby tissue from the EM sensor signal, and therewith from the associated model of volume air flow versus sensor signal. Use one of two related methods. Method 3A) Filter the raw EM sensor excitation signal using transform or circuit techniques to remove the acoustic pressure induced higher frequency noise, but preserve the needed low frequency excitation function shape information for model generated values of volume air flow and for subsequent feature vector formation. Method 3B) Use the tissue vibration signal from the EM sensor and the acoustic output (corrected for timing delays) to determine the backward acoustic transfer function. Divide the Fourier transforms of the vibration signal by that of the acoustic signal, and store the numerical (or curve fit) transfer function information in memory for recall as needed. Next, for each time frame, use the backward transfer function to calculate the glottal tissue vibration level associated with the measured output acoustic signal. Then subtract the backward transferred acoustic signal from the EM-sensor generated and processed signal, to obtain a "noise free" excitation function signal. This signal represents a backward traveling acoustic sound wave that induces

mechanical vibrations of glottal tissue and nearby air tract tissues in directions transverse to the air flow. This acoustic wave has little effect on the positions of the vocal fold edges, and thus it does not affect the actual volume air flow, U. However, certain EM sensors do measure this noise, and it shows up on the EM signal describing the excitation function (see Fig. 14B for an example). This noise level is found to be speaker specific. For high fidelity, speaker independent excitation function coding, such vibration signals mixed with the gross air flow values are undesirable.

4) Detect glottal tissue or nearby tract tissue motions that are transverse to the air flow axis and that are proportional to local pressure. Use, for example, a range gated EM sensor, optimized to measure the motions of pressure sensitive tissue, in directions transverse to the air flow axis. Calibrate using simultaneous signals from an EM sensor and from an air pressure sensor located near the pressure sensitive tissues. Use the EM sensor measured pressure, in each time frame, to determine air flow corrections in Equation (4). Correct those air flow values, due to post-glottal pressure variations that exceed the error-limits (user-defined) of the constant transglottal pressure approximation used in Equation (4).

5) Remove EM sensor measured noise on the glottal opening signal, by removing all signals not consistent with the mechanical equations of motion of the vocal folds (using known models such as those in Schroeter, J., Lara, J. N., and Sondhi, M. M.,"Speech Parameter Extraction Using a Vocal Tract/Cord Model," IEEE, 1987). Use EM sensors to measure and set the constants in the physiological model functions describing an individual's vocal fold motions, as described below in the section on physiological models. Use well known Kalman or other model based filtering techniques to filter signal contributions inconsistent with the model.

6) Insert an air flow sensor (and/or a pressure sensor) in the post glottal air tract and, using essentially simultaneous EM sensor signals, calibrate changes in transglottal air flow (and/or pressure) that are inconsistent with the model shown above in Equations 1-4, or for other models of air flow versus EM sensor signal. During training

sessions, obtain this data for the vocal tract configurations and for the frequencies where the effect is measured to be important for the application at hand. Then form a table lookup or a curve fit to associate each EM sensor signal value with a measured air flow value (and/or pressure value). During the actual speech application of the methods herein, obtain the EM sensor signal of glottal tissue motion. Associate the sensor signal with model values of uncorrected air flow or pressure, and then correct the air flow and/or pressure values as follows: 6A) Use the table of EM sensor versus pressure data to correct each post glottal or transglottal pressure estimate in the preferred model approach (e.g., Equations 1-4), or 6B) Use the table of EM sensor versus measured volume flow to directly correct each raw value of the air flow excitation function with a corrected value on a point by point basis. Describe the corrected pressure or air flow signals as amplitude versus time, or as Fourier amplitude and phase vs. frequency in transform space.

7) Change the model to make pressure the independent variable in the mathematical equations that describe the speech tract (for a circuit model example, see Figure 17B). Make volume air flow the dependent variable. The interchanging of voltage and current (i.e., pressure and volume air flow) between being the independent and the dependent variable in circuits and mathematical analogs is well known. See Figures 16, 17A, and 17B. Construct a table of EM sensor signal values versus measured pressure, for the range of vocal articulator conditions needed in the application as described in paragraph 6) and/or 4) above.

In summary, the algorithms obtains the excitation function, E(t), for each speech time frame, corrects it to the degree needed by the application by one of the above seven methods. The next, described below under the section on transfer functions, is to deconvolve it from the acoustic output to obtain the transfer function for the speech time frame and for the application. Experiments have validated methods, 1), 3A) and 6) above. Method 1) has been used to generate sufficiently accurate feature vectors for several speech recognition and speech synthesis applications. Method 3A) has been used to remove high frequency noise from the vocal fold area versus

time signal and method 6) has been used to calibrate an EM sensor against vocal tract air flow.

Formation of Voiced Feature Vectors:

The volume air flow function data provides, for the first time, a valuable description of the human voiced excitation function during each glottal open/close period of voiced speech. Most importantly, it enables the user to obtain the exact shape of the air flow vs. time and the duration of the vocal fold closure time (i.e., sometimes called glottal "zeros"). Figures 14A,B show annotated experimental data of measured glottal openings versus time. Typical triangular-like pulse shapes are seen. The sequence of individual pitch periods (i.e. single period speech time frames) are essentially all the same; thus a multi-time frame feature vector is easily formed. Secondly, this data shows a time offset between the acoustic signal and the EM sensor signal. This is caused primarily by the time of flight difference in timing between an EM signal reflected from the glottal tissues and the much slower acoustic signal which travels a longer path from the glottis, out the mouth/nose to the acoustic microphone. If timing corrections are needed, calibration procedures can be employed using laryngoscopes, air flow or pressure sensors, EM sensor calibration procedures, and/or accurate time measurements.

The glottal air flow (or pressure) amplitude vs. time can be used and coded in a variety of ways. They include describing the real time amplitude versus time interval, taking the appropriate transform, and/or approximating the shape by appropriate functions such as polynomials, a one-half sine cycle, piece-wise polynomials such as a triangle, and other similar functions. One example of coding the excitation function for minimum bandwidth transmission is to measure and store the excitation function feature vector as the parameters of a triangular open/close glottal area function versus time. It is described by the pitch period, the fraction of the period the folds are open (using the convention that the glottis opens at the start of the pitch period), and the location in the period of the peak opening and its magnitude (the peak amplitude is normalized). This simple description is more accurate than many presently used excitation functions and, for this example, is

described by only 3 numbers of 4 to 8 bits each. Furthermore, if several periods are measured to be "constant" in pitch period duration and acoustic output, the sequence of such periods can be represented by the single period plus one more number describing the number of periods of constant acoustic output, defining a multiple pitch period time frame.

A more complex excitation function feature-vector formation approach is to take the Fourier transform of the volume air flow vs. time over one or more glottal periods during which the acoustic speech units are constant and repetitive. An example is a long spoken /ah/ phoneme that is vocalized over a 0.3 sec duration. The feature vector and time frame are formed to describe the excitation function over a 0.3 sec time duration of substantially constant speech. For example, the user can record the frequency location of the highest amplitude signal (which is the first harmonic) that is the pitch or pitch period. In addition, the user can record the fractional amplitude levels of the higher harmonics compared to the fundamental harmonic, the phase deviation of the higher harmonics from the fundamental, and the bandwidth of the fundamental. Higher harmonic (e.g., where $n\ \omega_0 > 10\ \omega_0$) amplitude intensity relationships to the fundamental can be modeled knowing the mechanics of the vocal folds or by recording the experimentally measured rate per octave of fall, usually -12db.

Multi-time-frame feature vectors are formed by testing for constant or slowly changing waveform signals over several voiced speech periods. Constant means the acoustic and excitation amplitudes vs. time are nearly identical from one frame to the next, with nearly identical being defined as the amplitude in each time interval being within a chosen fractional value of a defined standard. This degree of constancy to a standard can be easily defined by the user ahead of time and automatically employed. The capability of this method to define constancy over one or more speech time frames using automated procedures is valuable because it enables economy of computing and increased accuracy of the functional descriptions. The reason is that one needs to only do one computation, using several speech frames with more repetitive amplitude data in contrast to performing a separate computation over each and every speech frame.

In addition, the user can define a slowly changing function that describes the change in volume-air-flow (or pressure) excitation over several speech time frame intervals. Examples of decreasing pitch periods occur during syllable emphasis or during a question. A feature vector can be formed over a time frame of several pitch periods, which contains the basic excitation function constant from a single period time frame together with one or two numbers that describe the functional change over the defined time frames. Fig. 14B shows the slight change in constancy of a voiced excitation over several speech periods as the speaker says the phoneme /ah/. This procedure also provides a means of defining a feature vector based upon deviations from the voiced excitation function of an average speaker or from the stored feature vectors of a specific speaker. In this case, the feature vector contains the deviations from average values, not the absolute values. This can be done in real time or Fourier space, or using mixed techniques.

Figures 9A,B, 10A,B and 11A show data taken by a male speaker saying the phoneme /ah/ for 36 consecutive glottal open/close speech periods, and derived speech functions. These figures illustrate the amplitude vs. time signals from the acoustic microphone and a glottal EM sensor (Figs. 9A,B), the Fourier power spectrum of each set of sensor signals (Figs. 10A,B), and the speaker's vocal tract transfer function (Fig. 11A) obtained by deconvolving the data in Fig. 10B from 10A. Using the procedures described below, a feature vector was formed over a time frame of 300 ms, in which the descriptors of the excitation function were taken from the Fourier transformed glottal function in Fig. 10B. The feature vector formation process is illustrated in Figures 12A,B. Experiments using data, as illustrated in Figs. 9A,B, show that the computation time to obtain pitch values, using the methods herein, is five times faster than by using conventional acoustic processing techniques, and the pitch values are more accurate than conventional acoustic-based techniques by over 20%.

Master Timing:

The method of measuring the glottal open-close cycle allows the user to define master timing intervals or "frames" for the automation of many speech technology applications. In particular, it

allows the vocalized excitation function periods to be the master timing intervals for the definition of time frames in the processing steps described herein. This approach allows the user to define the beginning and end of a glottal open/close cycle, and it provides a well defined method to join the information from one such cycle to the next cycle. It enables the concatenation of the information obtained in one speech time frame to be joined to that obtained in the next speech time frame. Figures 14A,B are illustrations of master timing, where each time frame is defined as one glottal cycle (i.e., pitch period), and the associated information is measured and labeled. Fig. 15B shows a sequence of single pitch period speech time frames for the spoken word "LAZY", and Fig. 15A shows the simultaneously measured acoustic information. One can define absolute pitch, the time frame duration, and characterize the timing information and store it as part of the speech frame feature vector which describes the acoustic speech unit spoken during the time frame. The cases when unvoiced speech segments occur are discussed in the section on unvoiced excitation.

The use of the glottal time period as the master timing signal allows the user to define time frames consisting of several glottal periods. See Figs. 14B and 15B for illustrations. The user sets algorithmic criteria to define "constancy" of the speech features being measured in order to determine how long the voiced speech time frame lasts. Then the algorithm measures how many pitch periods were used during which the "constancy" of feature values existed which are being used to describe the acoustic speech unit just sounded by the speaker. In the example above, the algorithm decided that 300 ms of constant sounding of the phoneme /ah/ took place. In this example, one of the "constancy" variables measured, and determined to be sufficiently constant, was the repetition frequency of the 36 glottal open/close cycles. The algorithm then defined a feature vector that described the time frame duration, the excitation function amplitude versus time for one period, and other information as shown in Figs. 12 A,B. Such a feature vector describes the acoustic speech unit, to the degree needed by the user, for the entire duration of the time frame. Because of the multiple glottal periods, the algorithm can average information obtained over

one or several of the included pitch periods, it can measure small period to period feature coefficient variations (e.g., pitch period variations) from the average which are useful for speaker identification, and it can use Fourier (or other) transforms to determine the voiced excitation function over as many or as few pitch period intervals as desired (or as many as the Fourier transform algorithm allows).

In the case that the speech changes from voiced to unvoiced, the last glottal open/close period of the voiced speech sequence has no "next" glottal cycle to use to define its end of period. In one approach, the algorithm continually tests the length of each glottal closed-time in each time frame for excessive length (e.g. 20% longer than the preceding glottal period closure-time). If the period is texted to be too long, the algorithm terminates the period and assigns, for example, a glottal-closure time-duration equal to the fractional closure time of the glottal function measured in the preceding time frames.

This method of defining constancy of speech over several glottal periods saves computation time and storage space in the computing processors and memories needed for many applications. It also allows the acoustic speech (and other instrument outputs) to be timed in a speech time frame along with other feature vector information obtained using the above timing procedures. For many examples herein, the feature vector is timed by the start time of the first glottal period provided by a master clock in the processor and its duration is defined by the number of constant glottal periods. This process automatically results in significant speech compression coding because feature vectors defining periods of constancy, as defined herein, can be shortened to one glottal period, plus a single number describing the number of glottal periods used.

The procedures above allow the definition of a time frame and the formation of feature vectors in which some of the coefficient values are slowly and predictably changing over a sequence of glottal pitch periods. An algorithm can define a time frame, over which slow changes in feature values (i.e., coefficients) take place, as follows. It measures the change in the coefficient value (e.g., pitch period) and fits the sequence of changes over several glottal cycles to a predefined

model. If the values do not fit the model, then a time frame with one or more slowly changing feature vector coefficients is not formed. If the coefficient values change too much, beyond the allowed range, an end of the time frame is defined. For example, a linear decrease in pitch period by 0.5 ms per cycle might be measured over 5 sequential glottal cycles, as a speaker "inflects" the pitch during the sounding of a single phoneme, when a question is asked. The algorithm also examines the other feature vector coefficients being measured during the time frame, but not being examined for slow change, to be certain that they remain sufficiently constant as demanded by the algorithmic definition of a speech time frame.

An example of such timing is shown in Fig. 14B where the first speech frame time period is 8.5 ms, the second is 8.0 ms, the third is 8.0 ms. A master clock in the processor times the onset of the first frame to be at 3.5 ms, the second at 12.0 ms, the third at 20.5 ms. The pitch deviations, referenced to the first frame, are -0.5 ms/ frame referenced to the first frame. The constant time offset between the fast closure of the glottal folds and the onset of the acoustic set is 0.7 ms, which is caused primarily by the differences in the distances and the speeds of signal travel between the EM sensor signal and the later arriving acoustic signal at the microphone. Such a time offset value does not influence the Fourier deconvolution process, as used in these examples. Another offset number is defined as the acoustic/EM frame-offset (or AEM number) by this method. It has value for recording the acoustic signal timing with respect to the EM signal timing. It allows the user to define the zero time of the acoustic signal with respect to the speech frame start. This characterization has value for speech to lip synchronization applications where sound to lip or other facial motion synchronization is required.

An example of a multiple pitch period time frame can be defined using measured data shown in Fig. 14A for the phoneme /ah/. By testing that the three measured pitch period changes referenced to the first pitch period, are 0.5 ms or less, and defining that a 0.5 ms change is constant enough for an application then a multi-period time frame can be formed. The other information in the sequence of feature vectors

must also be tested, and assuming it is also constant enough (for example the acoustic information in Fig. 14A is constant enough), a multi time frame can be formed into one feature vector describing a time frame 3 glottal periods long. One particular method for defining the pitch of the 3-pitch period vector is to use the average pitch period over the three frames, which is 8.16 ms; the average pitch deviation can also be measured and stored. Also in this example, the speaker was slowly raising his pitch (i.e., the pitch period shortened by 0.5 ms) as commonly occurs when stressing the end of a sound. This change can also be identified by the algorithm and stored if desired.

Using these methods the user can associate with each feature vector the start, duration, and stop times of the time frame using a continuous timing clock in the processor. The user can also store the absolute and relative timing information of the EM sensor information relative to other information (e.g., the acoustic signal) as part of each feature vector. Such timing information can be used to subsequently reconstruct the acoustic and other information in the proper speech order from the information contained in each single or multiple frame vector. In cases where the acoustic signal from the combination of the excitation and transfer function is known to last longer than a single glottal period speech frame, the transfer function information obtained allows the user to identify the part of the acoustic waveform that extends into the next speech period. The user is able to use such acoustic signal amplitude information in the time frame under consideration as needed.

The methods herein allow the user to conduct additional simultaneous measurements of speech organ conditions with instruments other than EM sensors. The methods herein allow the user to define "simultaneity" using the master timing information procedures described above for such measurements as video, film, electrical skin potential, magnetic-coil organ-motion detectors, magnetic resonance images, ultrasonic wave propagation, or other techniques. The methods herein allow synchronization, and incorporation into the feature vector for each time frame as desired, of such instrumentation output.

Unvoiced Excitation:

Using the general methods described above for voiced speech, one can determine the unvoiced excitation functions of the speaker and define unvoiced transfer functions, as well as speech frame timing and feature vector coefficient values. The method uses the algorithmic techniques for voiced/unvoiced detection that are described in the copending patent application Ser. No.08/597,596, filed 2/6/96. This algorithm uses EM sensors, especially the vocal fold EM sensor signals, to determine that acoustic speech is occurring without glottal open/close motions. Speech without glottal cycling is unvocalized speech.

The user selects (automatically or manually) an appropriate modified "white noise" excitation function that has been validated by listeners, by analysis, or derived using deconvolved functions as described herein. Such noise functions are characterized by their power spectrum per unit frequency interval. For excitation function feature vector formation, either a pattern (or curve fit) of the spectrum can be stored, or a numerical value can be stored which represents one of the small number of unvoiced excitation spectra needed for the application. Other EM sensors can be used (if available) to determine the source of the vocal tract constriction (e.g., the tongue tip, lips, back of tongue, glottis) and a modified white-noise excitation source appropriate to the air turbulence source, with proper noise spectrum, can be chosen. Once the source is defined, the chosen excitation function transform is divided into the acoustic output transform to obtain the transform of the transfer function of the vocal tract. The process to obtain the transfer function is identical to methods described above for generation of voiced transfer functions.

Unvoiced Speech Time Frames and Feature Vectors:

Unvoiced excitation functions can be obtained by using the methods described above in the section on processing units and algorithms to deconvolve the transfer function from the output signal to obtain the excitation function. The user first asks a speaker to speak phoneme sequences in a training session, using unvoiced phonemes, during which an acoustic signal is recorded. The user then uses general

knowledge of the speaker's acoustic tract, obtained from the literature or by using transfer functions, obtained by using voiced versions of the identically formed unvoiced phonemes. An example is to use the transform function from the vocalized phoneme /g/ to obtain the excitation function for the unvoiced phoneme /k/. The user performs a deconvolving operation to obtain the transfer function by removing the tract influence from the acoustic signal. The user then obtains the unvoiced excitation function used by a given individual in the measured speech frame. The user then stores the functional description for the specific individual, as a set of coefficients in an excitation function feature vector (i.e., to determine the noise generator spectrum), either using real time, transform, or mixed techniques. Typical uses of this and similar functions are for the deconvolving of acoustic output (during real time speech) to obtain a transfer function for complete feature vector formation, using processes as described in the section on feature vector formation. The full or partial feature vector for each unvoiced acoustic speech time frame is then available for the user chosen application.

The following three methods can be used for forming acoustic speech unit time frames when unvoiced speech is being sounded.

1) The user measures the time duration that an unvoiced excitation of acoustic speech units (e.g. phoneme or series of phonemes) is being sounded, during which no "significant" change in the spectral character occurs. This constancy definition for turbulence-induced sound is usually measured in frequency space where relative amplitude changes per predefined frequency intervals can be easily measured. For this method, "no significant change" is defined by first setting variation (i.e., constancy) limits within which the transform of signal levels must remain. Then during speech processing, each appropriate signal, such as the spectrum of acoustic output and other available EM-sensed organ-motion signals, are examined to determine if "change has occurred". A simple example of "change" is to use an EM-sensed start of glottal open/close motion to signal the algorithm that a transition to vocalized speech has occurred, and thus unvoiced speech has stopped being the

sole excitation. The duration of each unvoiced time frame is defined to be the total time of constant unvoiced speech, until a sufficient change in the acoustic or EM sensor signal occurs to signal the algorithm that a new time frame is defined.

2) A default algorithm is defined to accumulate data as in 1) above for 50 ms (or other user chosen time), and to define a 50 ms long speech frame and associated feature vector if no change in the constancy of the feature vector coefficients has occurred. If acoustic speech or a sufficient organ condition change occurs before 50 ms has passed, then the frame is terminated and the elapsed time to the event is the time frame duration. Otherwise, when a time period of 50 ms has elapsed, the speech frame is terminated and defined to be 50ms in duration.

3) An average vocalized pitch period of the user, taken during a training session (or normal speech) using a series of voiced words and phrases, is used as the default timing period for the unvoiced speech segments. The unvoiced period can be a non-integer multiple of such an average-defined time frame duration.

A method of defining slowly varying unvoiced speech is to analyze the unvoiced acoustic spectra every 10 ms (or user chosen minimal sampling period) to determine the degree of change per sample time. If the changes in spectra are slow or of low amplitude, then the longer time scale spectral variations can be characterized by a few parameters that characterize slowly varying noise spectral weights, the shorter term changes can be modeled by a few "dither-rate" spectral composition parameters, and the overall on-off amplitude envelope by an on-rate and off-rate parameter. These values, carried with the fundamental noise spectral values, can be formed into a single feature vector that characterized a time frame describing a relatively long segment of unvoiced speech.

Combined Voiced and Unvoiced Speech:

A small number of speech sounds are generated by using both a voiced and unvoiced excitation function. An example is the word "lazy" (see Figure 15) which transitions from a voiced-vowel sound of the phoneme /e/ (i.e., the "a" in lazy), to the voiced /z/ which

includes an additional fricative excitation in the oral cavity, and the word finishes with an /i/ sound. In those cases where two excitation sources are in play, the following procedure is used. The voiced excitation is first measured and deconvolved from the acoustic signal. However, since the Fourier transform of the transfer function still contains wide band spectral-power caused by the modified white-noise of the unvoiced sources, it may be removed as needed. Three procedures are available to detect, process, and code such signals:

1) The transfer function is tested for a noise spectrum which has an abnormally high frequency pattern showing it is not caused by normal pole or zero transfer function filtering of the vocal tract. If noise is detected, its spectral character is used to select an unvoiced excitation function for storing in the feature vector. Using the identified source, then a second deconvolution of the transfer function is taken to remove the influence of the unvoiced excitation function. The feature vector is formed for the time period and it includes descriptions for two excitation functions as well as the twice deconvolved transfer function, acoustic data, prosody parameters, timing, and control numbers for the application at hand.

2) The voiced excitation function is measured using EM sensors, and is deconvolved from the acoustic signal. No special test is used to determine the unvoiced noise spectrum. The resulting transfer function is fit with a predetermined functional and the nonvoiced excitation function is incorporated as part of the fitting. The result may have a higher-than-normal high frequency background in amplitude vs. frequency space. The coefficients are stored in the feature vector for the speech time frame. This procedure is adequate for most applications except those where very high fidelity synthetic speech is required. A variant on this method is to purposefully incorporate a noise functional into the transfer functional that is used to obtain a numerical fit to the deconvolved numerical transfer function.

3) Use one or more additional EM sensors to detect the conditions of the vocal tract that may lead to a nonvoiced excitation. For example if EM sensors, measuring the tongue-position, indicate that the tongue body is closing the vocal tract against the palate behind the teeth,

the tongue is in a position to cause turbulent air flow. An example is the unvoiced sound /s/, which with voicing added, becomes a voiced-fricative sound /z/. By using knowledge of the voiced excitation from the glottal sensor and tongue location, the algorithm can select the correct transform and deconvolve it from the acoustic waveform transform and test for noise presence. The next step is to test the transform for the noise spectral shape. If present, remove it with a second transform as in 1) above. This provides an acoustic transfer function transform, together with excitation function coefficients for forming a feature vector. This method is valuable because the user may not need to test every speech frame for the voiced/unvoiced excitation conditions. Yet, when it occurs, the method accurately performs the characterization as it is needed.

<u>**Transfer Functions:**</u>

The excitation of the human vocal system is modified by the filtering properties of the vocal tract to produce output acoustic speech. The filtering properties are mostly linear and are understood (for the most part). They can be described by linear systems techniques, as long as the necessary data is available. Traditional all-acoustic procedures do not provide the needed data. The methods herein obtain the necessary data and process it into very accurate descriptions of the vocal system for the first time. In addition, the methods obtain the data rapidly, in real time, and describe the human transfer function by a small number of parameters (i.e., coefficients) for each speech tract configuration. Additionally, the methods herein describe aspects of the human vocal-tract transfer-function that are important for speech quality but that are not well understood by experts. They enable a description of rapidly changing vocal tract configurations associated with rapidly articulated speech. They can obtain both the resonances and the antiresonances of the speech tract filter function (i.e., the poles and zeros of the transfer function), and information in real time, in frequency-space, or using combined descriptions. They also make possible the description of non-linear response as well as linear response transfer functions, because the output as a result of input can be stored in tabular form.

<u>ARMA technique</u>:

The transfer function can be obtained using a pole-zero approximation technique called the ARMA (auto regressive-moving average) technique, which makes use of time series or Z transform procedures well known to the signal processing community. This method of speech coding, using ARMA, provides a very convenient, well defined mathematical technique to obtain the coefficients defining a transfer function. Such a transfer function describes the vocal tract for each defined speech time frame. The ARMA deconvolving method includes obtaining substantially simultaneously, EM sensor and acoustic information, including amplitude, phase, intensity, and timing. In particular, the method provides a feature vector describing the transfer function by using the poles and zeros of the pole-zero ARMA description for the speech time interval frame or frames being coded. Alternatively, one forms a feature vector describing the transfer function by using, as feature vector coefficients, the a and b values of the a/b value description. (For signal processing references see Oppenheim and Schafer "Discrete-Time Digital Signal Processing" Prentice-Hall 1984", or Peled and Liu, "Digital Signal Processing: Theory, Design, and Implementation" Wiley, 1976). The poles and zeros describe the locations of the vocal tract filter resonances and antiresonances. The methods herein provide fundamental information, for the first time, describing the transmission "zero" frequencies of the vocal tract. The pole and zero values, or alternatively the a and b values, give the relative contributions of the resonances and antiresonances of the human vocal tract to the output acoustic signal.

For example, an ARMA functional was used to select 10 zeros and 14 poles for the sound /ah/, by using a least squares fitting routine. Figs. 9A,B show first the measured simultaneous acoustic and vocal fold EM sensor signal. The vocal tract Fourier transform is obtained by first taking the acoustic transform, see Fig. 10A, and dividing it by the EM sensor glottal function transform, shown in Fig. 10B. The deconvolved result is described by a series of complex numbers, or amplitude and phase values. The transform amplitude versus frequency, for the time frame, is shown in Fig. 11A. A 10 zero, 14 pole

ARMA model was then fit to the resulting vocal-tract transfer-function. Fig. 11A shows the numerical fit of the data to the ARMA functional, and Fig. 12B shows the pole/zero values that fit the phoneme /ah/. Fig. 11B shows a similar fit to the phoneme /ae/.

A feature vector for the speech time frame, during which a male speaker said the sound /ah/, was formed by obtaining, processing, and storing the information needed to characterize the acoustic speech unit to the accuracy desired, and is shown in Figures 12A,B. The feature vector includes several types of information. It includes the type of transfer function used. It indicates whether the segment includes a single phoneme or multiple phonemes. It provides phoneme transition information, for example the degree of isolation from previous and following phonemes. It describes the total time of constant excitation and counts the number of frames in the total vector. It also includes a description of the excitation function using the Fourier amplitudes and phases of the fundamental and the harmonics. This feature vector uses a predefined ARMA functional based upon the pole and zero value coefficients shown in Fig. 12B. An alternative functional description for the ARMA approach could have used the "a" and "b" coefficients, shown in Fig. 12C. Normalization and quantization methods were not used to form the feature vector in Figure 12A.

For the first time the user can capture the essence of an individual speaker's voice to a very high accuracy, because the user of the methods herein is able to approximate the actual data to a very high degree of accuracy. The approximation process is conducted consistent with the information content in the original signals and consistent with the numerical methods used in the functional definition processes. The ARMA method described here allows the user to capture filtering, resonance and antiresonance, and feedback effects that have not been previously available to the speech community, but which are known to be necessary to capture human voices (e.g. especially women's and children's voices). Examples of structures that characterize an individual's voice are known to be associated with complex nasal structures, non-circular vocal tubes, tissue compliance effects, mucous

layers, feedback effects on membranes, and other acoustic physiological interactions.

Predefined and Constrained ARMA Functionals:

Once the ARMA functional representation is obtained to the satisfaction of the user (depending upon the speech application and market), the user can "freeze" the functional representation for use for all work in a particular application environment. For example, the 14 pole, 10 zero ARMA functional may be the best one to use for a general purpose speech recognition application; but a different functional or set of functionals (e.g., 20 poles and 10 zeros for voiced nonnasal sounds, or 8 poles and 10 zeros for closed mouth voiced nasals) might be better functional choices for another user's application. The user could choose to take data from many speakers of a similar type (e.g. adult male American English speakers) using a fixed functional, but with differing pole and zero locations and with differing a and b coefficients reflecting their physiological differences. For many applications, the user will choose to average the defining parameters for the functionals and use them in a reference feature vector for code book formation. The user could also decide to use a training or adaptive process by which the system measures key physiological parameters (e.g. total tract length) for each speaker, and uses these data to pre-define and constrain the primary poles and zeros for each speaker. Using processes defined below, these pole-zero values can be normalized to those obtained from a reference set of speakers.

The user can use the procedures, and through experimentation define "More-Important" and "Less-important" poles and zeros in the ARMA expansion (where importance is a function of the application and value). "More-important" values are fixed by the well known major tract dimensions (e.g., glottal to lips dimension and mouth length and area) which are easily identified in the transfer function data and fit by automatic means. These values may vary from individual to individual, but their pole and zero positions are easily measured using the procedures herein. "Less-important" refers to those pole or zero terms whose contributions to the numerical fitting of the data are small. (One can use the "a" and "b" coefficients similarly).

These "less important" (higher order) poles and zeros are associated with the individual qualities of each speaker, and thus their values are very dependent upon the special qualities of an individual's tissues, tract shapes, sinus structures, and similar physiology that are very difficult to directly measure. This method of dividing the coefficients describing the transfer function into "More-Important" and "Less-important" categories makes it possible to generate feature vectors that are simplified and useful for communications. For example, only the "More-Important" values need to be sent each frame and the "Less-important" values can be sent only once, and used to complete the feature vector at the receiver end of a vocoder to improve the speaker's idiosyncratic qualities. Similarly, only the "More Important" values need be sent, thereby minimizing the bandwidth needed for transmission.

Finally one can associate (develop the mapping) from the ARMA parameters to the parameters that are associated with physiological, circuit analog, or other models which may be easier to use for real time computations than the ARMA approach. These other procedures are described below. This procedure is known to work because the ARMA "b" coefficients represent the signals reflected from the pre-defined vocal tract segments, and the "a" coefficients can be associated with zeros of known and unknown resonances. The signal reflections from vocal tract segments can be related to reflections from circuit mesh segments, or physiological tract segments. The engineering procedures for making such transformations from reflections to circuit parameters are well known.

The constrained functional method makes use of speaker training to limit the values of the poles and zeros (or a and b coefficients) to be near previously measured values. These constraint conditions are obtained by initial training using phoneme sounds that are well known to be associated with known vocal tract conditions. Adaptive training using a speech recognizer can also be employed to identify phonemes to be used for the definition phase. Physiological parameters are extracted from the transfer functions of phonemes chosen for their close association with certain tract configurations. An

example is to use the voiced phoneme /eh/ which is a single tube tract from the glottis to the lips; its primary transfer function resonance location provides a physiological measure of the speaker's tube length. With the total length known from the sound /eh/, the sound /ah/ allows the user to automatically define the division of the total tube length into the two sections from the glottis to the tongue hump. A series of these procedures are used to determine the dimensions of the vocal tract. Once these values are known, they can be used to constrain the ARMA functional variables during each natural speech frame. This process leads to faster convergence of the method to obtain the feature vector coefficients, because only a small number of fitting parameters need be tested against the data from each speech frame. In addition, these physiological parameters contribute numerical dimensions describing each individual speaker's vocal tract which contributes to speaker identification.

<u>ARMA feature vector difference coding</u>:

The difference feature vector method of coding allows one to define a feature vector by storing differences in each feature vector coefficient, c_n. The differences are formed by subtracting the value measured and obtained in the frame under consideration from the same coefficient formed during a previous time frame. For minimum bandwidth coding (also speech compression) the comparison is usually to values obtained during an earlier frame in the same segment when the algorithm noted that one or several important coefficients stopped changing. For the application of comparing a user's speech to that of a reference speaker or speakers, the reference feature vectors are obtained from a codebook using an additional recognition step. This method of forming such difference feature vectors is valuable because it automatically identifies those coefficients, c_n , that have not changed from a present frame to a reference frame. Consequently the information needed to be transmitted or stored is reduced.

If the reference values are predefined for the application, a complete difference vector can be formed (except for those control and other non-changing coefficients). Examples of reference speaker's feature vectors are those that describe the acoustic speech units of an

American English male speaker, an American English woman speaker, or child, or a foreign speaker with a typical dialect when speaking American English. The identification of the type of speaker makes possible the selection of appropriate functionals for more effectively coding the user's speech. Similarly, the speaker's own coefficients can be measured at an earlier time and stored as a reference set for identification applications at a later time. However if an application such as minimum information generation, is being used, a "mixed" algorithmic approach can be chosen by the user, wherein a complete, new coefficient value is stored in the vector location in the first time frame it appears, and then in the following sequence of time frames that show no change or slow change of the coefficient, only a zero or small change value is stored.

The procedure of forming difference vectors is conducted on each speech frame. The processor automatically compares the obtained feature vector to the defined reference vector, subtracts the differences for each coefficient and stores the differences as a new difference feature vector. This procedure requires that the reference procedure be previously defined for the acoustic speech unit vector under consideration.

The simplest method subtracts the appropriate feature vector coefficients obtained in the present time frame t_i from those in a frame measured at an earlier time t_{i-q}. Each coefficient difference, Δc_n, is placed in the "n" location of the difference vector for time frame t_i.

$$\Delta c_n(i,q) = c_n(t_i) - c_n(t_{i-q})$$

In the special case that q=1, and if the coefficient difference Δc_n is less than a predefined value, a zero value can be assigned to this nth coefficient in the difference feature vector, e.g., $\Delta c_n(i,i-1) = 0$. Similarly, differences of vector coefficients from values stored in vectors from any preceding or following time frame, e.g. t_{i-q} for q<i as well as for q>i, are straightforward to generate, and, if needed, can be tested for difference value levels.

For reconstruction, the identically zero value tells a subsequent application algorithm to look to the first preceding time frame, e.g. t_f with f<i-q, in which the examined feature vector

coefficient, $c_n(t_f)$, is non-zero. Upon finding a non-zero value, the coefficient value $\Delta c_n(t_f)$ is substituted for $c_n(t_i)$ for use by the subsequent application. If the application algorithm needs absolute values of the c_n's, then the full value feature vector must be reconstructed by using the predefined decisions for first finding the reference coefficient value. When using the difference vectors, the algorithm adds the difference coefficient value from the difference vector to the reference coefficient value to generate the coefficient $c_n(t_i)$, in the frame under consideration.

In the application where the measured coefficient vector values must be compared to those of a reference vector coefficient, two approaches are possible. Either known speech segments are spoken by the speaker for which references have been previously recorded, or a speech recognition step must be employed to first identify the feature vector under consideration and to then find the associated reference feature vector. In this way the subtraction of coefficients can occur and difference coefficients can be used to form a difference vector describing the acoustic speech unit or units in the time frame.

This method of differences is valuable to minimize the amount of information needed for storage or for transmission because many of the vector coefficients will be zero. Consequently they will take less storage space, computation time, and transmission bandwidth. The absolute feature vector for the speaker can be reconstructed at a later time as long as a definition standard for the coefficient zeros (or other no-change symbols) is known or is transmitted along with the feature vector, e.g. the identical zero code described above. An example of importance to telephony is to first store a standard speaker's feature vector values, for all phonemes and other acoustic units needed in the application. These data are placed in both the recognizer processor and in the synthesizer processor codebooks. Then, whenever an acoustic speech unit is to be transmitted over the medium, only the unit symbol and the deviations of the user speaker from the reference speaker need be transmitted. Upon synthesis, the average speaker coefficients stored in the receiver, plus the deviation coefficients, form more accurate vectors for reconstructing the text symbol into speech.

Another important application is that this automatic method of determining deviations from standard speakers saying known sounds, enables algorithms to self adapt the system. When certain reference sounds are pronounced and certain difference vector coefficients exceed a predetermined level, the algorithm can trigger an automatic "normalization" of the speaker's feature vector to that of a reference speaker for more accurate recognition or other applications. Conversely, if the differences become too large, over a short time period, the algorithm could signal appropriate persons that a personnel change in the user of the system has occurred.

<u>Electrical Analog of the Acoustic System</u>:

The excitation function and the transfer function may be approximated as defined above, using well known electrical analogs of the acoustic system. See Flanagan 1965 for an early, but thorough description. Figure 16 shows a simplified electrical analog of the human acoustic system showing an excitation function, a vocal tract transfer function impedance, and a free air impedance. By fitting the circuit parameters of the equivalent electrical circuit, each time frame, to the measured excitation function and transfer function data, automated algorithms can determine the "circuit" parameter values. The advantage of this approach is that the relatively small number of types of human vocal tract resonator conditions (10 to 20) can each be modeled by a set of circuit elements -- with only the specific parameter values to be determined from the speech information each time frame.

For example, Figs. 17A,B show an electrical analog of a straight tube human acoustic system with electrical analog values, e.g., the L, C, R's, which represent the acoustic coefficients of a single tube system which is used for the acoustic speech sound /ae/. Using the deconvolving approach illustrated in Fig. 5 and using the transfer function values in Fig. 11B, the impedance values shown in Fig. 16 and the circuit values shown in Figs. 17A,B can be determined for the sound /ae/ using algorithms to fit the circuit values to the transfer function data. Feature vector coefficients can be defined by using the electrical-analog transfer function as the functional representation and by using the electric circuit parameters to represent the transfer function. The

parameters are easily fit to the well defined transfer functions because the methods herein show how to separate the excitation source from the vocal tract transfer function in real time for each speech time segment. In addition to the methodology of forming a feature vector, the electrical analog circuit parameter values are useful in describing the physiological vocal tract values because the L's represent air masses, the R's and G's represent acoustic resistance and conductance, and the C's represent air volumes. These physiological parameters can also be used as feature vector coefficients.

For the single mesh circuit in Fig. 17A, the air volume velocity transfer function between glottal and mouth is given by the following expression, which includes radiation load:

$$\frac{U_m}{U_g} = \frac{\cosh(\gamma_r L)}{\cosh(\gamma + \gamma_r)L}$$

where γ and γ_r are related to the mesh circuit parameters as given in Figure 17A and are defined as:

$$\gamma = \sqrt{(G + j\omega C)(R + j\omega L)}, \qquad \gamma_r = \frac{1}{L} tanh^{-1}\left(\frac{A_t}{A_m}\left[\frac{(ka)^2}{2} + j\frac{8ka}{3\pi}\right]\right)$$

A_t and A_m are the area of the throat and mouth opening respectively, and k is the wave number of the sound, and a is the radius of the mouth opening. For the case of a simple tube such that $A_t = A_m$ (i.e., the case of equal glottal and mouth area) the poles of the transfer function are given by:

$$S_n = F(a, L)\left[-(\alpha c + \frac{a^2\omega^2}{2Lc}) \pm j\frac{(2n+1)\pi c}{2L}\right]$$

where

$$F(a, L) = \frac{3\pi L}{3\pi L + 8a}$$

$$n = 0, 1, 2, \ldots \qquad (1)$$

The physical parameters in Eq. (1) are: L, the vocal track length; a, the mouth opening radius; and α, the vocal tract wall resistance. Typical

numbers are: F(a,L) ~ 0.94; a ~ $5.2e^{-4}$ cm^{-1}; and the speed of sound c = $3.5e^4$ cm/sec. The low order poles can be determined. They can be used to constrain the physiological variables using the equations below. The three physical parameters can be estimated from measurements of the first two pole locations on the S-plane. They are r_0, r_1, ω_0, and ω_1, the corresponding real and imaginary parts of the first two poles of the transfer function. Then the three physical parameters can be determined from the following relations:

$$a = \sqrt{\frac{2\pi c^2 |r_0 - r_1|}{(\omega_1 - \omega_0)(\omega_1^2 - \omega_0^2)}} \tag{2}$$

$$L = \frac{1}{3\pi}\left(\frac{3\pi^2 c^2}{(\omega_1 - \omega_0)} - 8a\right). \tag{3}$$

and

$$\alpha = \frac{1}{c}\left(\frac{r_1(3\pi L + 8a)}{3\pi L} - \frac{a^2\omega_1^2}{2Lc}\right) \tag{4}$$

__Physiological Parameters__:

The methods used for obtaining the information described above can be used to generate a feature vector using the physiological parameters of the human speaker vocal tract as the coefficients to describe the acoustic speech unit spoken during the speech time frame. The transfer function parameters used to define the ARMA models, the electrical analog model values, and those obtained from real time techniques described herein, define physiological parameters such as tract length, mouth cavity length, sinus volume, mouth volume, pharynx dimensions, and air passage wall compliance. In addition to the physiological parameters, the feature vectors would contain, for example, the excitation function information, the timing information, and other control information.

One can then use this physiological information as coefficients of a feature vector, or they can be include in the ARMA or other transfer functional forms to constrain the coefficient values. For example, once one knows the tract length from glottis to lips by saying the phoneme /ae/, one knows the basic resonance of the speaker's vocal tract and it serves as a constraint on data analysis by defining the lowest frequency formant for the speaker.

An example of the data that is available using the methods herein is to use the pole zero numerical fit to the transfer function data for the sound /ae/ shown in Fig. 11B. The lowest formant pole, f_1, is at 516 Hz, and using the simple expression, neglecting the radiation term, one finds the vocal tract length:

$$L=\frac{c}{4f_1}=\frac{3.5e^4 cm/sec}{4*516}\cong 17cm$$

Similarly, the pole zero data for the sound /ah/ in Fig. 11A provides the data for the glottis to tongue hump plus tongue hump to lip data.

An important application of the physiological values is that they provide a method to normalize each unique speaker's transfer function to that of an appropriate average speaker. In this manner, each formant value, obtained through deconvolving methods herein, can be transferred to a new value by using measured physiological values and instant reference values.

Another important use of physiological parameters is to measure the glottal and vocal fold mechanical properties as phonemes are voiced. The EM sensor that measures the glottal structure motion, enables the user to constrain the mechanical values of the glottal mechanisms. These values include opening amplitudes, spring and mass constants from the pitch, and damping, and compliance from sympathetic tissue vibration due to backward propagating acoustic waves (i.e., low pressure acoustic waves). Special phonemes are chosen for calibration purposes, such as those with the low post glottal pressure (e.g., open tube phonemes) like /uh/ or /ah/.

The differences in physiological conditions and in excitation functions for well known phonemes allow an automatic identification of several attributes of the speaker. This can be used for

identification purposes as discussed above, but also can be used to automatically select the best types of transfer functional forms to be used to fit each user's physiology. Examples are to identify gross features of the speaker vocal tract dimension, e.g. an adult male, an adult female, a child, and other variations well known to the speech practitioner.

Speech Coding:

The purpose of recording and coding EM sensor and acoustic information is to use it for specific user defined applications. The methods herein include processes to define the characterizing parameters for a variety of physical, engineering, and mathematical models that are valuable and useful for all EM sensor/acoustic based speech technologies. They include processing procedures, which include time frame definition, coefficient averaging, normalization, quantization, and functional fitting to convert the EM sensor/acoustic data to form feature vectors. These methods are mostly linear procedures, but are not limited to linear techniques. Examples of nonlinear procedures include, but are not limited to, taking the logarithm of the acoustic data or the transfer function to reflect the human hearing function, or to compress the frequency scale of the transformed data in a linear or nonlinear way (e.g., "Mel" or "Bark" scales) before the functional fitting techniques are used. Such processing depends upon the application. Feature vectors for appropriate time frames can be formed by fitting linear or nonlinear functional coefficients to the processed data, and such feature vectors can be stored into code books, memories, and/or similar recording media.

The vast amount of data generated by the methods herein, measured over a wide frequency range for every speech frame, enable the definition of the coefficients used to fix the functional forms into functions that fit the data. For example, the EM sensor data shown in Figs. 9B and 10B for the phoneme /ah/ was generated at 2 MHz and the simultaneous acoustic data (Figs. 9A and 10A) were digitized at 11 kHz (using 16 bits). This provides 250 EM data points per acoustic point, which are averaged to match the accuracy of the 16 bit acoustic data. In each nominal 10 ms speech frame, this leads to 80 averaged data points per EM sensor and 80 acoustic data points to define a set of functional

coefficients. In principle between 80 and 160 unknown coefficients can be determined. However experts skilled in the art of fitting functional forms to data know how to use such large data sets to define a smaller number of coefficients associated with simpler model-based functional-forms. In particular, the flexibility of the techniques described herein make it possible to design the EM and acoustic data collection systems that work well over a very wide range of data accuracy and detail.

Single- and Multi-Time-Frame Feature Vectors

Using the methods herein the user can describe the excitation function, the transfer function, the speech time frame parameters, acoustic parameters, prosodic information such as pitch or amplitude envelope shapes (obtained during one or a series of time frames), and control information (e.g. types of transfer functionals and frame clock times). The user can easily assemble this information into a feature vector for each speech time frame. These individual time-frame feature-vectors can be joined together to describe concatenated vectors describing several acoustic speech units occurring over two or more time frames (e.g. diphoneme or triphoneme descriptors). Such a multi-time-frame feature-vector can be considered as being a "vector of vectors". These multi-time-frame feature vectors can be constructed for all phonemes, diphonemes, triphonemes, multiphonemes (e.g. whole words and phrases) in the language of choice. They can be stored in a data base (e.g., library or code book) for rapid search and retrieval, for comparison to measured multi-time-frame feature-vectors, and for synthetic speech and other applications. The capacity to form a feature vector describing the variations in speech units over many time frames is valuable because the time varying patterns of the sequences of the individual vector coefficients are captured by the corresponding sequence of speech frames. This approach is especially valuable for storing diphone and triphone information, and for using Hidden Markov Speech Recognition statistics on defined sequences of many (e.g., 10 or more) acoustic speech units.

A specific example of describing a long duration, multi-phoneme speech segment is to "sample" and define the feature coefficients every time a change in coefficient condition is detected, as

described above for single time frame vector formation. At each time of condition change, t_i, a feature vector of p coefficient values, $c_n(t_i)$, where n=1 to p, is obtained (see Fig. 12A). This procedure produces a sequence of sets of feature vector coefficients that are obtained at the specific times of change noted by the values $t_1, t_2, \ldots, t_i, \ldots, t_k$. For example, the time values, t_1, denote the start time of the speech frame. However the t_i's can also denote a sequential frame number noting the frame position in a sequence of frames. Because the time frame duration is usually included in the feature vector as the pitch period or the number of pitch periods (or other notational forms), the total time taken by a frame or a sequence of frames (i.e., comprising a speech segment) can be reconstructed. For example, below is a set of sequences of p coefficients $c_1(t_i), c_2(t_i), c_3(t_i), \ldots c_p(t)$ for each start time $t_i = t_1, t_2, \ldots, t_k$.

$$c_1(t_1), c_2(t_1), c_3(t_1), \ldots c_p(t_1), \;\; c_1(t_2), c_2(t_2), c_3(t_2, \ldots c_p(t_2), \ldots,$$
$$c_1(t_k), c_2(t_k), c_3(t_k), \ldots c_p(t_k)$$

This method describes an adaptive procedure for capturing the essential speech articulator information throughout a speech segment, without requiring a frame definition every 10 ms as many acoustic (CASR) recognition systems do. These patterns of coefficient sets form a multi-time-frame feature vector that describes an entire speech segment that begins at time t_1 and ends at time t_k + (last frame duration time). Such vectors, which can include pause times (i.e., silence phonemes) are very unique for each speaker. They time compress the coded speech information, and they store all of the information needed for the application by choice of "change" condition definitions, and by choice of sensors, accuracies, and other considerations described herein.

Normalization and Quantization:

Normalization:

The methods described herein can code any type of acoustic speech unit, including coarticulated or incompletely-articulated speech units. The coding methods provide very high quality characterization of each spoken phoneme for each spoken speech segment, but if the articulation of the user-speaker is different from those speakers whose acoustic speech units, or sequences of speech units, were used to generate the reference code book, then the recognition or other process

loses some accuracy. The unique ability of the methods herein to characterize the physiological and neuro-muscular formation of each speakers articulators makes it possible to normalize each unique speaker's transfer function to that of an appropriate reference speaker. These normalization methods reduce the variability of the feature vectors formed during each time frame by normalizing the feature vector coefficients (or sequence of units) to those of a reference speaker or speakers.

During a training session, the user speaks a series of speech units or speech unit sequences into systems like those shown in Figs. 3A,B. A group of feature vectors are selected by asking the user to speak a desired vocabulary, or by using speech recognition during natural speech to select the desired vocabulary. The coefficients of each speech vector, for every selected speech time frame, are compared to the feature vector coefficients from the same reference words generated by a reference speaker at an earlier time. In this way, all the feature vectors for the acoustic speech units needed in the reference vocabulary are measured and placed in a reference codebook at an earlier time.

The process begins as the algorithm compares each measured vector coefficient, c_n, to that of the reference speaker each time frame. If it differs by a predefined level (e.g., a user chosen 20% value), then either the coefficient in the reference codebook or the one in the speaker's feature vector is to be changed. This process of normalization is carried out for each speech time frame, using one of the three following methods:

1) Codebook Modification: All feature vectors listed in the codebook and which relate to the tested acoustic speech units in the limited vocabularies, have their coefficients changed to be those of the speaker specific feature vector. Also included is a process for altering those multi-phone sound-unit sequences in the code book, which contain individual word sounds in need of correction. Acoustic sound units that are correctable, e.g. phonemes, diphonemes, and triphonemes, contain coefficients that are often associated with "misarticulated" phonemes. The specific coefficients of the multiphone feature vectors are altered to reflect the idiosyncratic articulation of the associated single

speech unit as determined during training. For example if the speaker misarticulates the sound /th/ as in "the", then all diphonemes, triphonemes, etc. that have /th/ in them such as /th/ /a/ /t/ in the word "that" are corrected to the speaker's feature vector. Similarly, multiphoneme units can be spoken, compared, and changed in the codebook as defined by this algorithmic prescription. This procedure leads to the construction of a speaker specific codebook.

2) Key Sound-Sequence Modification: During the training session, the speaker articulates special acoustic sound sequences that are known to be poorly pronounced by speakers of the language. The acoustic sound unit sequences are measured using methods herein and feature vectors are formed. The measured feature vector coefficients for these multi-unit articulator conditions are stored in place of similar feature vector coefficients in the predefined codebook locations. This provides a partially "individualized" multi-phoneme codebook.

3) Method of Extremes: The speaker says a series of training acoustic speech units that require the speaker to use his articulators in their extreme positions or rates (e.g., highest to lowest position, fastest to slowest rate, front-most to back-most position). By finding the feature vector representations for these extremes, using both direct EM sensor methods and the deconvolving methods, one obtains two extreme limits on the coefficients describing each feature vector coefficient. The extreme coefficient values, for each coefficient c_n are represented by ${}_{min}c_n$ and ${}_{max}c_n$. These two extreme values can be used, for example, to represent the longest and shortest vocal fold periods and the largest and smallest of each transfer function coefficient for acoustic speech units. Other values, such as the average value of the extremes, ${}_{ave}c_n = ({}_{min}c_n + {}_{max}c_n)/2$ for each coefficient in the feature vector coefficient location, c_n, can also be obtained. These special values are stored in a separate, but "parallel" codebook that contains the "user extremes", user averages, and other useful values that correspond to each user coefficient, c_n, that will be used in the formation of normalized feature vectors for the application.

The next step in the method of extremes is to generate the needed reference speaker extremes, averages, and other useful values as

well. Each reference speaker (or speakers) is asked to articulate the set of identical sound units for the training cycle of the speaker being normalized. Next, the sets of reference coefficient extremes (as well as other information such as averages) are associated with each coefficient c_n for each acoustic sound unit in the separate, but "parallel" codebook. An example of other useful values are those that represent special articulator conditions that define intermediate articulator coefficient values. These are valuable to aid in non-linear or guided interpolation procedures.

During normal usage of these methods, when the speaker speaks any sound unit, a time frame is defined and a feature vector is generated. Each measured coefficient, $_{meas}c_n$, of this feature vector is compared to the maximum ($_{max}c_n$) and minimum ($_{min}c_n$) range of the speaker's coefficient extension for this coefficient c_n.

The fraction of distance, f_n, of the measured coefficient between the two extremes of the speakers range is calculated, using as an example a linear approach as illustrated in Figure 18:

$$f_n = {}_{meas}c_n \,/\, ({}_{max}c_n - {}_{min}c_n)$$

The coefficient $_{meas}c_n$ is then replaced with the coefficient $_{normal}c_n$ as follows, using the minimum and maximum ranges of the reference speaker.

$$_{normal}c_n = {}^{ref}{}_{min}c_n + f_n * ({}^{ref}{}_{max}c_n - {}^{ref}{}_{min}c_n)$$

In this equation, f_n contains the information from the user's own measured c_n value, and from the "parallel" code book of extremes containing the user's and the reference speaker's extreme values (and other useful values) associated with each feature vector coefficient, c_n. In this way the fraction of the user's articulator coefficient range is mapped to that fraction of the reference speaker's range.

This procedure is very easy to implement because the acoustic speech unit in each time frame is characterized with a relatively small number of coefficient values that require normalization (e.g., a sub-set of the coefficients c_1 through c_p in Fig. 12A). It is well known that other interpolation techniques for f_n can be used as desired, besides the linear one described above. In addition, it is clear that control coefficients such as timing and phoneme symbols whose numerical

values are contained in one or more of each feature vector's coefficient values are not normalized as described above.

The above normalization methods enable the user to correct for incomplete articulation because the feature vector coefficients associated with incomplete articulator positioning are normalized to the correct coefficient values articulated and recorded by reference speakers. In addition, coarticulation is corrected by normalization of multi-speech-frame vectors that describe diphonemes, triphonemes, and similar acoustic units where coarticulation most commonly occurs. It is important to note that the extreme values (i.e., target values) for each phoneme in a multiphone sequence as determined from a reference speaker or speaker group will be different than for individual phonemes or other primitive speech units from the same reference persons. That is, the speech organ articulators do not reach the same extreme values of c_n associated with isolated phonemes when they speak the same phonemes imbedded in di-, tri-, or higher order multiphones.

The voiced pitch value of an individual speaker is an important coefficient that can be normalized to those of the reference speaker or speakers as described above. The procedure is to normalize the appropriate excitation feature vector coefficient, c_n, which represents the pitch value (i.e., the reciprocal of the pitch period) of the speaker for the voiced speech frame under consideration. The pitch value extremes for both the speaker and the reference code book contain maximum pitch, minimum pitch, and intermediate pitch values as needed (e.g., a pitch value for each of the major vowel groups). The normalization of the excitation function pitch-value coefficient proceeds as described above for generalized coefficients.

Since a person's physiological tension level, as well as external stress or health factors, can change a user's pitch, rate of speech, and degree of articulation, it is important that they be corrected as often as the application allows. Daily pitch normalization is available using the first words a user speaks to turn on the machine or to "log in". Adaptive updating, using easily recognized vowels can be used to correct the maximum and minimum levels, as well as the intermediate normalization values as shown in Figure 18A. As the day progresses,

and the user tires or becomes stressed, adaptive correction based on automatically recognized acoustic speech units can be used.

Quantization of Feature Vector Coefficients:

It is known from speech research that the vocal articulators must move or change some condition a minimal amount for a perceived change in the speech sound to occur. (See references by Stevens, "Quantal Nature of Speech: Evidence from Articulatory - Acoustic Data" in "Human Communication--A Unified View" eds. David & Denes, McGraw Hill, 1972.) Thus changes in the values of these feature coefficients and pitch values that do not cause a perceived difference in the application (e.g., recognition or synthesis) can be grouped together in a "band" of constant value. As a consequence, during training and synthesis experiments, the user can determine the bands of coefficient values, using a reference speaker or speaker groups, over which no perceptible speech changes are detectable for the application at hand. Once these bands of constant speech perception are determined, for each applicable feature vector coefficient, including excitation function coefficients, the measured coefficient values, c_n, can be quantized into the value of the band. As speech takes place, each measured feature vector coefficient is first normalized, and then "quantized" or "binned" into one of only a few "distinguishable" values. Figure 18B shows such a procedure based upon the normalization procedures described above and illustrated in Figure 18A.

The algorithm proceeds as follows. First, the feature vector coefficients are measured for each speech time frame. Second, each coefficient is normalized to a reference speaker's value for the coefficient as shown in Figure 18A. Third, each normalized coefficient value is quantized into one value that represents a band of constant acceptability over which the coefficient can vary in value, but produce no discernible change as defined by the user. Thereby a continuum of coefficients can be mapped into only a few values, representing a few bands. The band coefficient value is usually chosen as the central value of the band. If the normalized coefficient, ${}_{normal}c_n$, is in the range spanned by the second band of the reference speaker's discernible bands, then the measured value ${}_{meas}c_n$ is mapped first to ${}_{normal}c_n$, then into the

quantized value ${}^{2}c_n''$. The double accent '' means the coefficient is quantized and the superscript 2 refers to the second of the bands spanning the total range of the normalized feature vector coefficients ${}_{normal}c_n$.

If the user wishes, quantized band values obtained during reference generation and during use can be further normalized. For example each of the n bands can be associated with a fractional value ranging from 0 to 1 (or over another range of the user's choice) for numerical convenience. For example, it may be desirable to quantize pitch rate into 3 values, such as 1, 2, and 3, representing low, middle, and high frequency pitch of any speaker, and to not use absolute pitch frequencies such as, for example, 70 Hz and 150 Hz, or similar physically meaningful values. This method of normalizing quantized values is valuable because it removes all apparatus and speaker specific values, and it enhances table lookup speed and accuracy.

<u>**Real Time Measuring, Recording, and Deconvolving:**</u> The methods described herein permit the user to select the appropriate techniques for sensing, processing, and storing the information with an almost arbitrary degree of linearity, dynamic range, and sampling bandwidth for the desired application. They can be used in a variety of configurations depending upon the costs, the value of the data, and the need for portability and convenience. Because of the flexibility of these methods to meet the needs of a wide variety of applications they are very valuable.

The method of using real time information to relate excitation-source signal-features to related acoustic-output signal-features, is valuable for obtaining physiological information for several applications. For example, these procedures can be incorporated into a training sequence when a user first begins to use systems based upon the methods herein. By requesting the user to speak a known series of phonemes, the algorithm can be automatically adapted to the user (or by using speech recognizers that recognize key phonemes from which the desired timing information can be extracted). For example, the methods allow the determination of the acoustic tube lengths of an individual as known phonemes are spoken. The phoneme /ae/ is known to be

caused primarily by a voiced, single tube resonance from glottis to lips to the microphone. The time it takes for an excitation signal to travel the length and appear as an acoustic signal can be measured and used to determine parameters used in the vocal models of an individual's speech tract. (see Figs. 14A,B for an example of time duration). The knowledge of the length permits faster numerical model fitting, because one of the major tract filtering properties is constrained. It is also valuable in speaker identification, by providing a physiological measurement that contributes to the definition of a unique speaker.

Similarly, in other speech tract configurations, such as a nasal /m/, the sound travels from the glottis through the nasal passage, as well as into the closed mouth resonator. The sum of the two signals exits the nose to the microphone. An acoustic echo (canceling certain frequencies in the speech output) will be caused by the closed mouth resonator. Other phonemes are caused by similar combinations of tubes and resonators. The glottal excitation travels differing paths, have differing time delays. The real time methods described herein enable the measurement of these other tract dimensions as well.

This method provides for deconvolving, in real time, the excitation source from the acoustic output to obtain useful vocal tract information. The dimensions and other characteristic values of the user's vocal tract segments, obtained for each speech segment, can be used to form a feature vector to describe the vocal tract for subsequent applications. Experiments have provided physiological values for the phonemes /ah/ and /ae/.

Applications:

Speech Compression: The methods provide a natural and physically well described basis for speech time compression. The methods defined above for difference feature vector formation, for multi-time-frame feature-vector formation, for multiple glottal period time frames, for slowly varying feature vector time-frames, and for unvoiced time frame determination show algorithmic descriptions of accurately coding speech segments using much less time than real time spoken speech. Simple extensions of these methods show how to collapse both the silence PLU e.g., pause speech segments) to one vector and relatively

long unvoiced speech segments to one vector. These methods enable one to collapse time segments of essentially constant speech into one time frame and one representative (i.e., compressed) feature vector. The compressed vector contains only a few additional coefficients that describe how to "uncollapse" the speech back to real time as needed. Additional compression can be attained using grammatical and syntax rules that remove redundancy of sound patterns, such as a "u" always following a "q" in American English. These simplified patterns can be undone during speech synthesis, during reconstruction of transmitted speech symbols, or from speech stored in memory.

Speaker Identification: The methods of feature vector formation herein enable a user to compare a feature vector from one or several speech segments to the same speech segments as spoken by a reference speaker, and stored in a codebook for the purposes of speaker identification. The coding and timing methods for this purpose can be performed automatically, by defining the feature vector over each time frame or sequence of time frames. The identification operation can be conducted using the feature vectors from isolated time frames or using multi-phoneme time segments. The user is able to make identifying comparisons using previously agreed upon speech segments (e.g., names or PIN numbers) presented to a user by the system for his vocal repetition. Alternatively, speech recognition can be used to extract key speech segments from natural speech. The identified feature vector patterns (i.e., multi-time frame feature vectors) are compared to those in the reference codebook.

In addition to the frame by frame comparisons against reference frames described directly above, additional information on the average pitch and the pitch variations of the user, the physiological parameters of the user's vocal organs, and the EM wave reflection strength from the user (tests water and tissue composition) are available. These parameters are obtained from initial sound requests to the user by the system and are initially obtained as the user "logs in". They are then used for comparison against values known, by the system, to represent the true speaker.

The identification process uses a measurement algorithm that compares the distance of the measured feature vector coefficients from those stored in the codebook each time segment. At a normal speaker's rate of speaking 5 to 10 phonemes per second, a twenty to thirty phoneme sequence, with time spacing and prosody values, can be obtained within a few seconds. For very sophisticated recognition as much as a few minutes of speech may be required; and for very high value work, continuous recognition may be employed using speech recognition for continuous key pattern identification and verification of the speaker throughout the use period. During the sampling time, statistical algorithms process the data and obtain the probability of correct identification.

In addition to the acoustic and EM sensor patterns, physical parameters of the user can be obtained using the methods herein. The physiology of the vocal organs such as sizes, positions, normal positions (e.g. normal pitch), and tissue compliances can be obtained. Also the quality of articulation of each acoustic sound unit, as well as the rates of formation are obtained. Each speaker's unique articulation qualities are exaggerated when combinations of rapidly spoken sounds such as diphonemes or triphonemes, etc. are measured and compared to previously stored data. The methods herein describe how such multiphone feature vectors are formed, measures of distance formed, and measures are used for comparison. The organ dimension, articulation positions, and their time patterns of motion in conjunction with acoustic speech information, taken over a sequence of acoustic speech sounds, are very idiosyncratic to each speaker of any language.

This method makes possible the use of the feature vector coefficients to define a distance metric between the user's characteristics and those defined when the validated speaker spoke the same acoustic unit from which the vectors were formed and stored in a pre-defined library. One example measurement process is to obtain the distance between all the measured and stored vector coefficients (control and other special coefficients excepted):

$$\Delta C_n(t_i) = {}_{meas}C_n(t_i) - {}_{ref}C_n(t_i)$$

for all time frames denoted by the time of the frame, t_i. The algorithm then takes the square root of the sum of the squares of all the coefficient differences, $\Delta c_n(t_i)$, for all speech time frames in the sound sequence. If the measure is less than a pre-defined value, based upon previous experiments by the user, the user speaker is accepted as validated. This example method is a uniform distance metric applied equally to all appropriate coefficients. Other methods which use non-uniform coefficient weighting methods, non-linear measure processes, and which use differing statistical testing are well known.

Other applications use similar comparison procedures that are made between the speaker and reference libraries of vectors with coefficients obtained from averaged (or other types of reference speakers) to determine the physiological or linguistic type of speaker. For example a male American English speaker, female American English speaker, child, or foreign speaker with a specific dialect can be identified for various purposes.

Language Identification: The patterns of feature vectors vs. time (i.e., multi-time frame feature vectors) are very indicative of the language being spoken by the speaker. A method to determine the language being spoken by a speaker is as follows. It uses the procedures described above for speaker identification, except that a separate normalized (and quantized if need be) language codebook is previously formed for every language in the set of languages for use in the application. As the user speaks known test sounds, or by using real time recognition techniques to extract test sounds from the natural speech, the algorithm forms feature vectors for each speech period using the individual glottal period feature vectors as the basis. The vectors can be normalized and/or quantized as needed. The algorithm then forms these basic patterns into more complex patterns and it searches each one of the several language code books for the measured patterns. The patterns are chosen to contain the unique identifying sound patterns of each language. The algorithm then uses the statistics of appearance times of multi-time frame feature vectors, of specific vocal articulator positioning represented by specific or small groups of feature vector coefficients (especially glottal pitch patterns), and it searches for the appearance of

those unique sound patterns associated only with a given language. Several methods of measuring multi-component vector distances, are available to test for the best fit and are described above in the section on speaker identification. When a best fit of the speech segments to one of the language codebooks is found, the language of speech is identified and the probability values of the recognition are available as needed.

Speech Recognition:

The methods described herein make possible the identification of all spoken acoustic speech units in any given language in a new and powerful way. This new type of speech recognition is based upon using the feature vectors defined above using processed information from the excitation function, the deconvolved transfer function, simultaneously recorded and processed acoustic information, and the timing information. The feature vectors are more accurate than those based upon acoustic techniques alone. The reason is that they are directly tied to the phonemic formation of sound segments. They are more accurate than other approaches because both poles and zeros can be accurately modeled, the pitch can be accurately and rapidly measured, and the feature vector coefficients can be readily normalized and quantized, removing speaker variability. The vectors describe the condition of a speech unit with sufficient information, including redundancy and model constraints, that the phoneme (or other acoustic speech units) can be defined, with very high probability, in an automated fashion for each speech time frame. An identification results when the measured and processed phoneme feature vectors from a speech segment are associated with a stored reference vector containing the symbol or symbols of the acoustic speech unit. The acoustic speech unit identification results in a recognized symbol (e.g., a letter, pictogram, series of letters, or other symbol). Once the speech segment's identification symbols are available, they can be automatically coded to ASCII (or other computer coding) or to telephony codes for transmitting letters, pictograms, or text symbols over communications channels. Such procedures to convert recognized acoustic speech symbols into "technological codes" are known to practitioners of communication technologies.

Methods for normalizing tract feature vectors and excitation functions, for time independent acoustic description, for normalizing rates (i.e., time warping), for dealing with coarticulation, incomplete articulation, and phoneme transitions can be used to simplify the variability of measured patterns of speech information between individuals and by the same individual at different times. These make possible more rapid and accurate code-book "look-up" of the correct acoustic-speech -unit symbol.

Training, Table Lookup and Table Generation:

A training process is used by algorithms described herein to ask a speaker (or speakers) to articulate a known vocabulary of speech segments into a system similar to one shown, for example, in Figs. 3A or 3B, 8, or 20. The segments can range in complexity from simple phonemes to continuous natural speech. The training process enables one to build up known associations of measured feature vectors with symbols for known acoustic speech units by using the instruments shown in the representative systems and the methods described herein. The system designer can select the appropriate processing algorithms from those described herein, including normalization, quantization, labeling and other necessary operations to form and store the feature vectors for each trained sound segment into a code book location or library locations (i.e., a data base). These code-book data-sets serve as references for most of the applications described herein. Methods of associating a measured speech feature vector with a similarly formed set of vectors in a code book make use of well known procedures for data base searches. Such procedures allow the algorithm to rapidly find the locations in the data base where the measured vector matches stored vectors. Procedures are described and to rapidly calculate vector distances to determine the best match, and to determine probabilities of association. Accurately formed feature vectors, normalized and quantized, allow for very rapid data base searches.

An EM/Acoustic Template Matching Model for Speech Recognition:

The feature vectors can be used for phonetic template (i.e., pattern) matching and associated acoustic speech unit identification. Each acoustic speech unit symbol is uniquely associated with a specific

articulator configuration (i.e., a phonetic articulator pattern). The formed vectors, which describe these patterns, are then compared against the library data and an identification is made using the "distance" from the code book feature vectors, and using logical operations, such as "on" or "off" for the glottal motions. In the case of speech segments with multi-phonemes, similar methods of measuring vector distances can be used. One procedure is to use the square root of the sum of the squares of all relevant vector coefficient differences. (Control coefficient distances are not used). When the distance is within a value defined by the user, an identification is defined and the related probability based upon the distance measure can be attached to the identification unit as desired. The use of a logical test operation is well known. Well defined normalization and quantization techniques for feature vectors make for well defined code book comparisons because the vectors can be instrument and speaker independent. An additional advantage is that individual-speaker rates of phoneme sequence articulation can be normalized and time aligned speech frames can be produced.

An EM/Acoustic Hidden Markov Model for Speech Recognition:

The methods of forming speech unit feature vectors by deconvolving the EM sensor measurement of the excitation function from the acoustic output can be used to form vectors of data from sequences of speech frames representing sequences of phonemes. They describe the coding of many sequential acoustical units, e.g., sequences of phonemes, diphones and other multi-phones. Such vectors are especially useful for the purposes of identifying symbols for natural spoken speech using an EM/Acoustic Hidden Markov Model (HMM) method. Many human speech segments consist of many phonemes run together, and are therefore many acoustic units long before word-breaks occur. Sequences of single speech frame feature vectors as well as one or more multiple speech frame feature vectors can be treated as patterns of numerical values that can be tested against combinations of the pre-stored patterns of the limited reference feature vector data set. HMM statistical techniques can associate these measured and formed sequences of feature vectors with test patterns constructed, as needed by the

algorithm, from only a limited number of feature vectors in a code book. Typical code books contain pre-recorded and processed feature vectors for 50 PLUs and 1000 to 2000 diphones.

An EM Sensor/Acoustic HMM allows the user to statistically identify a phoneme or a pattern of phonemes by comparing the probability of observing such a series of feature vectors representing known words or phrases. This procedure requires a learning phase, as is well known in the art for the acoustic vector HMM approach, to build up the test patterns of combinations of feature vectors for the words in the vocabulary being used. The methods herein make the HMM method of speech recognition very valuable, because the data is so accurate and well defined. The methods herein provide very accurate procedures to rationally identify feature vectors by deconvolving, normalizing, quantizing, time aligning, and modeling the recorded information. The algorithm then forms a sequence (i.e., matrix) of as many feature vectors as needed for the specific EM/Acoustic HMM in use. As a consequence most of the ambiguity of individual speaker variations is removed and the patterns of speech units have little variability from speaker to speaker making HMM a very accurate identification technique.

<u>An EM/Acoustic Neural Network Method of Speech Recognition:</u>

Neural network algorithms are useful for associating a pattern described by a feature vector with a symbolic representation of one or more acoustic speech units. This method uses the training period method to cause the adjustable parameters within neural network algorithms to be associated with the EM/Acoustic input feature vectors. Because these are speaker independent and instrumentation independent), the vectors defined during speech by a user as well as by reference groups of speakers during codebook generation have little variance for the same acoustic speech unit. The associating of the real-time, input feature-vectors is conducted using well known neural network algorithms (e.g., back propagation using two or more layers) to associate each input with a known acoustic speech unit, e.g., phonemes, words or other speech units. For the procedures herein, each feature vector may be 150 coefficients in length, which when taken three time

frames at a time, require nearly 450 inputs to the neural network. (control and similar feature vector coefficients are not used as inputs). Once trained, off line using a computation process of needed power, the network algorithm can be loaded into the user's processor to provide a rapid association from an input feature vector to an unambiguous output speech unit. (see for example Papcun et al., J.Acoust. Soc. Am. 92, pt. 1, p. 688 (Aug. 1992) for "micro beam" x-ray detection of speech organ motions for an approach well known to practitioners of neural network applications). Because of the unique association of a speech sound symbol with vocal articulator positions, as represented by the feature vector coefficients, an accurate identification of the symbol associated with each feature vector can be made.

<u>A Method of EM/Acoustic Joint Probability Speech Recognition:</u>

Recognition using the method of joint probability can produce increased speech recognition accuracy. It is based upon jointly using the deconvolving approaches together with conventional speech recognition (i.e., CASR) information, and using pure EM sensor based recognition information (i.e., NASR).

Step 1: The user chooses a conventional acoustic (CASR) system to examine an acoustic speech unit or speech unit series (e.g., phoneme series). The CASR system selects one or more identifications (e.g. phoneme symbols such as /ah/) which meet the criteria of identification. A first set of all such identified units, with probabilities of identification exceeding a user-chosen level (e.g., 80%), are formed.

Step 2: The deconvolving process, plus other information as described herein, is used to form a feature vector. One of the statistical techniques (e.g., HMM, phonetic template, or neural networks) is used to identify the symbols for one or more acoustic speech units associated with the feature vector formed during the speech frame being examined. If the identification is within the predefined probability band, it is associated with the identified sound unit symbol (and its actual probability of identification is also recorded) and it is added to a second set of identified acoustic sound units. Other potential unit identifications from this step, with differing but acceptable probabilities of recognition, are included in the second set as well.

Step 3: The user select data from an EM sensor system in use, and generates a NASR feature vector each speech time frame. The NASR system estimates symbols for one or more acoustic speech units that meet the probability criteria of NASR identification procedures. A third set of symbols of identified acoustic speech units is formed, with attached probabilities of recognition.

Step 4: Steps 1, 2, and 3 are each repeated to generate probabilities of identification for those symbols identified in the other steps that were not found the first time through. That is, an identified unit from step 1 with probability (for example) greater than 80%, could have been un-recognized in step 2, because its probability was below a cutoff value. For the joining of probabilities each symbol from each step must have a probability of identification from the other 2 steps. In the second cycle through, if a symbol is not easily assigned a probability in any one of the procedural steps, it can be assigned a probability of zero.

Step 5: An algorithm joins the separate probabilities from step 1 and/or step 2, and/or step 3, in a fashion weighted by their probabilities to obtain the most likely recognized sound unit. One algorithm is to find the joined probability by taking the square root of the sums of the squares of the probabilities for the symbol obtained from each step 1, 2, and 3.

The important and valuable addition provided by the deconvolved feature vector data, and other procedures herein, is that it is a mixing of acoustic with EM sensor data which provides an additional degree of data correlation that is sufficiently different in a statistical measurement sense that the joint probability of the data described above will be better than if only one or two separate sets of data were used. This approach works well with one EM sensor and microphone, but is especially valuable when the user chooses to employ two or more EM sensors with an acoustic microphone. This approach also works very well with multiple sets of very precise, but often incomplete data.

An example of a two EM sensor system uses an EM glottal motion sensor and an under-jaw, upward-looking EM sensor. With these the sensors, the user obtains three data sets from: 1) a single EM

sensor feature vector describing the conditions for the jaw, tongue, and velum signals each time frame, 2) glottal motion data from an EM sensor measuring the excitation function and 3) acoustic microphone data. Probabilities of symbol identification, using the data three sets can be joined together naturally by a single software processing system using standard statistical algorithms. Each individual sensor, plus the deconvolving of 2) from 3), offers very unique and precise features that lead to a high probability for certain sets of symbols and a very low probability value for all other symbols. Using all three sets together, the algorithm form a very high probability of identification of a unique symbol. The user has the option with such a combined system to use each sensor and algorithm in its most economical and accurate way for the recognition application. This approach leads to economical computing, and rapid convergence to the identified sound unit.

<u>A Method of EM/Acoustic Exclusive Probability Speech Recognition:</u>

The method of exclusive probability uses methods of formation of three sets of feature vectors described above in steps 1 to 3 in the section on joint probability speech recognition. It uses a sequential procedure to statistically reject identifications made by any one of the three types of recognition systems. It uses logical tests to exclude (i.e., reject) symbols not meeting certain criteria.

Step 1: Use the CASR approach to identify the acoustic sound units for the speech time frame or frames under consideration, as long as the probability of symbol identification exceeds a user defined value, e.g. 80%. At this stage, the probability criteria is set to retain symbol identifications that may have similar probabilities of identification by the CASR data at hand. Subsequent steps are be used to eliminate ambiguous identifications from this step.

Step 2: Use the deconvolved feature vector set to reject those identified sound units from 1) that meet the probability criteria of definition (by CASR) but fall below the user-set levels of acceptable probabilities for identifications of symbols based upon the probability of identification using the feature vectors formed by the EM/Acoustic methods herein.

Step 3: Use one or more of the NASR EM sensor identification methods to check the probability of each remaining identified acoustic unit symbol from step 2. Identify those acoustic speech units that do not meet the probability criteria of the NASR system, and reject them. Leave the remaining, highly probable acoustic units and their probabilities of identification in the data set.

Step 4: Use a standard statistical algorithm to join the probabilities of those identified acoustic units that remain in the set, after Step 3. This leads to a small number of acoustic speech units, usually one, that meets the "exclusion" criteria of the sequence of three steps.

This process rapidly eliminates those ambiguous identifications, caused by insufficient data at each step. Symbols that have low probabilities of identification are rejected early in the process and thereby reduce computational processing later in the process. This process causes the one or few remaining acoustic speech unit symbols, which pass the three sequential sensor/algorithm tests, to have a very high probability of correct identification. This method can be applied to the data by permuting the order of techniques for identifying the feature vector. For example, the deconvolving technique might be used in Step 1, while the CASR technique could be used in step 2. The method of exclusion can also work with two rather than three identification steps. This method is very valuable for using partial information from auxiliary sensors or as "by-products" of the major sensors. It provides a more accurate identification of the acoustic sound unit than either an all acoustic system, or an all EM/acoustic feature vector system could accomplish without the additional information. For example, the presence of one or more fast tongue tip motions measured with a tongue EM sensor indicates that the acoustic unit identified by the deconvolving process must be a phoneme consistent with such tongue motion, e.g. in English /th/ as in "the", or a rolled /r/ as in "rosa" in Spanish or Italian. If the feature vector coefficient from step 3, for example, does not describe rapid tongue tip motion, the symbol identification is rejected.

If two speech units symbols remain, that have sufficiently high probabilities, both placed in a set with their associated probabilities. The user can choose to use only the highest probability unit or the system can automatically ask the speaker to repeat the sound or phrase if both probabilities are similar or below desired certainties. If no recognized symbol meets the probability criteria, then a signal can be sent to the control unit that the acoustic speech unit is ambiguous, and the identified acoustic units are shown in order of certainty with probabilities attached. The algorithm can be programmed to automatically ask the speaker to repeat for clarification under such circumstances.

Speech Synthesis:

The methods provide for the synthesis of high quality, idiosyncratic speech from stored EM sensor/acoustic data obtained from an individual speaker or from an averaged set of speakers. Individual speaker means any individual, ranging from a normal office dictation worker to a famous actor. The speech encoding process to be used for subsequent synthesis depends upon how the original feature vectors were coded and stored in a code book. The methods herein can be used to form a set of feature vectors optimized for speech synthesis. They may be based upon an average speaker or a particularly desirable speaker whose acoustic speech is quantified and stored in a codebook.

Step 1: Form a reference codebook by recording the acoustic speech units of a desirable speaker or group of speakers for each acoustic speech unit needed for the synthesis application of the user. Form feature vectors of all of the acoustic units that will be used based upon the procedures herein, and use the master timing techniques herein to define the beginning and end of these vectors.

Step 2: Use a commercial text-to-speech translator that identifies all of the required speech units (phonemes, diphones, triphones, punctuation rules, indicated intonation, etc.) from written text for the purpose of their retrieval.

Step 3: Use an automatic search and retrieval routine to associate the sound units from Step 2 with a code book location described in step 1.

Step 4: Select the feature vector to be used from the code book location described in step 3. The feature vector information, in addition to excitation function and transfer function, includes the timing of the sound units, the joining relations from frame to frame, and the prosody information.

Step 5: If phoneme to phoneme transitions are not called out by step 2, generate the transition acoustic sound units using one or more of the following: Two sequential voiced sound units are joined at the glottal closed times (i.e., the glottal zeros) of voiced speech frames, while unvoiced frames (or unvoiced-voiced frames) are joined at acoustic amplitude zeros. If transition rules are present that describe the rate of interpolation between voiced phoneme units, they are used to set the transition time frame durations and to interpolate excitation and transfer function coefficients that are modified by their relationship to another articulator condition in the preceding or following time frame. Another method of interpolation is to use diphoneme or triphoneme acoustic speech patterns, pre-stored in a code book, which are normalized to the proper intensity and speech period and which are placed, automatically between any two phonemes called for from step 2.

Step 6: Provide the prosody for the acoustic sounds generated during each speech time frame or combination of speech time frames. For example, use prosody rules to set the rate of sound level amplitude increase, period of constancy, or rate of amplitude decrease over several speech frames. Use prosody rules to set the pitch change from the beginning of the speech sequence to the end, as defined by phrasing and punctuation rules. Such prosody information is obtained from the text-to-speech converter, in step 2, and is used to alter the frame vectors as they are taken from the code book to meet the demands of the text being synthesized into speech.

Step 7: Convolve the excitation function and the transfer function, together with the intensity levels, and generate a digital output speech representation for the time frames of interest. This procedure can produce acoustic signals that extend into the next speech time frame. The signal from one frame can be joined to the acoustic signal (i.e., amplitude versus time) generated in the next frame by procedures of

adding wave amplitudes and then squaring (coherent addition) or by squaring amplitudes and adding to obtain intensities (incoherent procedure). Combinations of these approaches, with "dithering" or varying feature vector coefficients from frame to frame, may be employed to simulate the short term variations in human speech. This digital representation is converted to analog, via a D/A converter, and broadcast as desired.

Figure 19 shows data for the reconstructed acoustic speech unit /ah/, which experimentally produced a pleasing sound. The originally recorded acoustic data is shown by the points on the curve and the line is the reconstructed sound spectrum, formed according the steps 2 through 7 above. The sound /ah/ was manually chosen.

Methods to Alter Synthesized Speech:

The methods of coding and storing speech feature vectors can be used to alter the original coding to meet the speech synthesis objectives of the user. The methods described herein provide the user with well defined and automated procedures to effect the desired speech changes. For example, the original speech pitch can be changed to a desired value and the rate of delivery of acoustic speech units can be changed to a desired rate. In each speech feature vector, several coefficients describe the excitation function. By changing the duration of the excitation function, either in real time (for example by compressing or expanding the individual glottal triangular functional shape to take less time) or in transform space (by moving the transformed excitation amplitude values to higher or lower frequency bins), one can change the pitch to be higher or lower. These procedures increase the number of glottal open and close cycles per unit time, and then by convolving this higher (or lower) pitch excitation function with the unchanged vocal tract transfer functions for each newly defined speech time frame interval, one obtains a new higher (or lower) pitch voiced output. To implement prosody rules, that describe pitch change, the algorithm can cause a rate-of-change of pitch to occur during a segment of speech, containing several pitch periods. The algorithm slowly changes the excitation function pitch for each frame, from an initial pitch value to a slightly higher (or lower) one in the following frame. Also, the

algorithm can "dither" the glottal period duration for each constructed time frame to provide a more natural sounding synthesized speech.

These new methods provide a very important procedure for joining sequential excitation functions during their periods of glottal closure. In this manner, no abrupt changes (i.e., no signal derivative discontinuities) takes place in the real time acoustic output signal. In a similar fashion, the user can simply add (or subtract) extra time frames or extend a multiframe transfer function (i.e., with constant excitation function and transfer function, just more periods) to adjust the length of each speech unit. Using these methods, one can extend the time it takes to say something or speed up the speaking to finish words sooner, but maintain excellent quality speech using the basic, speech-frame "building blocks" provided by the methods herein.

An important application of these methods is to synchronize the rate of an actor's speech recorded in a sound studio, with his or her facial motions (e.g., lips) on video (and/or film) media. The obtaining of facial vocal motion requires the use of an EM sensor to record lip motions and a video image analyzer to track key facial motions (e.g., lips) on video or film media associated with known speech frame features obtained using the EM sensor information. Image analysis systems are commercially available that can follow patterns within a video or film image. The methods herein allow the user to synchronize the speech track by synthesizing new speech, at correct rates, to follow the facial motions in the sequence of images. The algorithms herein can alter the excitation function length by stretching or compressing the time frame, by adding or deleting additional frames, by shifting frames in time by adding or deleting silence phonemes, by introducing pauses, by keeping certain frame patterns constant and by stretching others, and in such a manner that the apparent speech is unchanged except that it matches the facial motions and/or other gestures of the speakers.

The user may also alter the transfer function of the speaker as desired. The user can modify the physiological parameters and construct a new transfer function using physiological or equivalent circuit models. Examples are lengthening the vocal tract, changing the

glottis to mouth diameter ratio, or increasing the size of the nasal cavity. The methods also allow almost arbitrary changes in transfer functional construction for amusement, for simulating animal sounds, for research, or for special "attention-grabbing" communication applications by "playing" with the coefficients and synthesizing the resulting speech. Once a modified transfer function is formed, as a consequence of altering the physiological models or by using empirically determined coefficients, the user then makes the corresponding changes in the code book. All feature vector coefficients in the code book that correspond to the altered transfer function are changed to make a new code book. The methods herein enable such automatic modifications because the several functionals described above for defining vocal tract transfer functions, e.g., the ARMA, equivalent circuit parameters, or physiological based functionals, are well determined and easily modified. For synthesizing the modified speech, the user proceeds according the speech synthesis steps described above. Each selected acoustic speech-unit, is associated with a feature vector that includes the modified transfer function information, the excitation, prosody, timing changes, and control information (including synchronization data).

Another method of altering the data stored in a code book that was derived from one person or from an average person is to substitute the excitation function coefficient descriptors in a given feature vector by those from a more desirable speaker. Similarly, one can exchange the transfer function, or the prosody pattern from an original speaker with those from a more desirable speaker. The user then performs, upon demand, the convolving of the excitation function with the transfer function to produce a new unit of sound output for the purposes of the user. For consistency, such changes must be performed on all relevant feature vector coefficients that are stored in the code book being used. For example, all excitation function coefficient descriptors in all feature vector coefficients must be changed according to the prescription if one person's glottal characteristics are substituted for another's. This is easy to do because all feature vector formats are known and their locations in memory are known; thus, algorithmic

procedures allow the user to alter a known set of codebook vectors and their specific coefficients.

These methods for altering and reconstituting speech make it possible to generate synthetic excitation functions and transfer functions that are very unusual. Methods of change include generating animal speech by using animal vocal system models, constructing physically impossible open-close glottal time functions or transfer functions, shifting pitch periods to create very high pitched voicing (e.g., dolphin speech at 100kHz), or changing the excitation functions in response to external stimulus such as to follow musical sounds or notes. That is, a poor singer could sing into systems similar to those herein, and a musically corrected voice would be synthesized and broadcast. Or an animal trainer could speak into a processor and have his speech sounds transformed to those frequency bands and patterns optimized for the animal being trained. These techniques can easily create physically unrealizable feature vectors, based upon exaggerated physiological parameters. The technique can also create feature vector alterations to obtain amusing sounds (e.g. chipmunk voices) or desirable prosody patterns. These special effects can be used for purposes of entertainment or research, or other specially desired effects can be easily created using the techniques. Since the coding methods are both fundamental and convenient to use, these methods are very useful and valuable.

Speech Telephony

Analysis-Synthesis Telephony -- Vocoding:

The methods of speech recognition and speech synthesis described herein provide a valuable and new method of speech coding and decoding for the purposes of real-time Analysis-Synthesis Telephony (i.e., Vocoding). It is particularly convenient to use the feature vector generating process because the speech segment feature vectors are in a form immediately usable for synthetic speech and for telephony transmission. One method of analysis-synthesis telephony (i.e., vocoding) starts with a speaker speaking into a microphone while an EM sensor measures glottal tissue motions. Figure 20 shows a view of a head with a cutaway of a vocoding telephony handset 90. Handset 90 holds three EM sensors 91, 92, 93 and an acoustic microphone 94. EM

sensors 91, 92, 93 are preferably micropower radars optimized for specific organ condition sensing, and direct EM waves toward and receive reflected EM waves from various speech organs. For example, sensor 93 is positioned for vocal fold and glottal motion measurements. Handset 90 also includes a transmitting and receiving unit 95, which is connected externally through wired or wireless connection 96. Transmitting and receiving unit 95 is connected to a control unit and master clock 97, which controls a speech coding processor, recognizer code book and memory unit 98 to which EM sensors 91, 92, 93 and microphone 94 are connected. Control unit 97 is also connected to a decoder processor, speech synthesizer, memory and code book unit 99, which is connected to a receiver loud speaker 100. Unit 99 and speaker 100 are mounted in an ear piece 101 of handset 90 so that the speaker 100 is positioned over the person's ear. Several system functions illustrated in Fig. 20 are similar to those shown in Figure 8.

The speech is analyzed by deconvolving the excitation function from the acoustic output, and feature vectors are formed describing each time frame of the speech output. The numerical coefficients of these feature vectors can be transmitted directly using standard telephony coding and transmission techniques. Alternatively, the speech sound unit can be speech recognized, and the symbols for the recognized unit (e.g. in ASCII or other well known code) can be transmitted. Additional control or speaker characterization information can be transmitted as desired. The methods for the formation of "difference feature vectors" and for the identification of "More Important" and "Less Important" transfer function coefficients are especially useful for telephony because their use reduces the bandwidth needed for sending coded voice information.

At the receiving end of the telephony link, the transmitted signal is reconstituted into speech. The synthesis procedure may use the transmitted feature vectors, it may synthesize new speech from transmitted speech symbols, and using its internal code books of stored feature vectors in a "text-to-speech" process. The user may choose a combined approach using partial speaker information to "personalize" the synthesized speech to the degree desired. Alternatively, the

receiver's controller may recognize incoming coded speech, and direct the recognized symbolic information to a local computer system for processing or storage purposes, to a fax system or printer to print the received symbols, or to an analog recording system for later use by the intended receiver.

The method of vocoding herein includes the process of attaching additional information to the transmitted speech information-packet for each speech frame. This additional information can be used by the receiver to perform speaker identification, to do speech alteration, to translate to a foreign language, to encrypt the data, or to minimize the bandwidth. The transmission of the feature vectors thus formed can occur in real time over transmission systems such as wire, optical fiber, acoustic (e.g., underwater communication) or over wireless systems. The method then includes synthesizing the feature vectors into acoustic speech representing the speaker, for the purposes of broadcasting the rendered acoustic sounds through the telephony receiver to the listener. The speech synthesis part of the vocoding system can be designed to use average speaker qualities, or it can be designed to transmit very high fidelity speaker-idiosyncratic speech. High fidelity transmission will use relatively higher bandwidth for the transmission of the more accurate description of the feature vector information, than the minimum possible, but it will require much less bandwidth than present high fidelity voice transmission. Conversely, minimum bandwidth systems remove all information about the speaker except for that needed to communicate minimal voice information.

When the speaker in a vocoding communication system becomes the listener, and the listener the speaker, the vocoding system works in the same fashion as described above except for the interchange of speaker to listener, and listener to speaker. In addition the process can operate in real time, which mean that the recognizing, coding, recognition (if needed), and synthesizing can take place while users are speaking or listening. Real time means that the time delay associated with coding, transmitting, and resynthesizing is short enough for the user to be satisfied with the processing delay. The computationally efficient methods of coding, storing, altering, and timing, which have

been described herein, make possible the needed rapid coding and synthesis. Elements of such a system have been demonstrated experimentally by coding several spoken basic speech sounds and acoustically synthesizing them using the coded information.

Minimal Bandwidth Transmission Coding:

Minimum transmission coding is made possible using the identification and coding procedures described herein. One method is to use the speech compression methods described above. Another is made possible when the speech recognition part of the system results in a word identification and/or the sending of minimal speaker idiosyncratic information. By using speech identification in a system, such as the one shown in Fig. 20, each acoustic speech unit is translated to a word character computer code (e.g. in ASCII) is then transmitted along with little or no speaker voice characterization information, for the purpose of minimizing the bandwidth of transmission. The symbol transmission technique is known to use 100 fold less transmission bandwidth than real time speech telephony. Thus the value of this transmission bandwidth compression technique is very high. The speech compression techniques described above using the coding procedures herein, is less effective at bandwidth minimization, but it is simpler to use, retains most of the speaker's speech qualities, and is calculated to use 10 fold less bandwidth than real time speech.

Reductions in bandwidth (i.e., bandwidth minimization) can be attained using many of the well known coding techniques in present communications, most of which are based upon the principle of only transmitting changes in information that are discernible to the user and they do not retransmit information every "frame". The "difference feature vector" method described above is very useful for this application. In addition, bandwidth minimization is further enhanced by using the minimum quality of speech characterization needed for the application. The methods for the characterization and reconstruction of speech are especially suitable for these procedures of bandwidth minimization, because these methods herein show how to measure and characterize the simplest units of speech possible. For example, partial information on the speaker's physiology can be sent to the receiver's

process and incorporated into the synthesis model for more personalized speech reconstruction. Once obtained, these speech "building" blocks of excitation and transfer function can be approximated and used in many ways. In particular, well defined decisions on the "change information" needed to update the next frame of speech, consistent with the user's needs, can be made before the information is sent off through the transmission medium. Because the coding and resynthesis techniques are so intimately and naturally linked, the initial coding for transmission and subsequent decoding and resynthesis is straightforward and economical. These methods are valuable because they provide important means to save valuable and expensive transmission bandwidth that reduce costs. Another valuable use of the method is to allow additional information, such as encryption "overhead" or speaker identification, to be transmitted along with the sound information on present fixed bandwidth systems.

Simultaneous Spoken Language Translation:

The methods herein for real time speech coding, recognition, and resynthesis in a vocoding system are valuable for real time speech translation from one language to another.

Step 1: The user speaks into a system such as shown in Figs. 8 and 20. The system codes each acoustic speech unit.

Step 2: The system recognizes the coded speech units and forms symbolic text of the letters, words, or other language units such as pictograms.

Step 3: The system uses a commercial language A to language B translation system, which takes the symbolic text of the recognized acoustic language units from Step 2 and translates them into symbol text for the language B.

Step 4: The system uses a commercial (or other) text to speech converter to convert the symbols in language B into feature vectors, together with prosody rules.

Step 5: The system synthesizes the translated symbols into acoustic speech in language B.

A variant on this method is, in step 2 above, to associate with each recognized word in the codebook, the associated foreign word.

Thus the translation step 3 and the text-to-speech in step 4 is avoided for simple translations applications. This language translation system can work in real time and be very compact. It can be packaged into a portable megaphone (e.g., Fig. 20 but with a translation unit and a megaphone attached) where the user speaks one language and another language comes out. For more complex and more accurate translation applications, it can be built into a stationary system as shown in Figure 8.

Presentation and Teaching:

This method of feature vector formation makes it possible to display the information received for each speech unit for feedback to the user. The display information can be graphical on a screen (e.g., images of the speaker's vocal tract), or the information can be sounded, printed, or transmitted to a user via tactile or electrical stimulation. The use of feature vectors based upon physiological parameters aid in the visual display of the sizes and positions of the vocal tract articulators of the speaker. These can be used for purposes of speech correction, real time speech assistance, and speech education because the information can be used to illustrate the problems with the positioning of the speaker's vocal organs for the attempted sounds. Conversely, the methods herein enable the illustration of the corrected vocal organ positioning for the desired sound, using reference codebooks of correct feature vectors. These procedures are very valuable for speech correction and for foreign language teaching. The capacity to recognize the user's speech and to communicate the characteristics of the speech back to a disabled user, in real time, is of great value to speech impaired persons. For example, a deaf speaker can receive feedback stimulus, via tactile or electrical signals to his skin or to his inner organs, on the quality of their articulation.

Conclusion

The invention includes a method of measuring and generating in an automatic manner an accurate speech excitation function of any speaker for one or several sequential speech time frame intervals. Simultaneously, the acoustic signal is measured and the excitation function is deconvolved from it, leading to a speech tract transfer function for one or several sequential speech time frame

intervals. The invention includes methods of accurately timing, coding these data into feature vectors, and storing the information into code books.

There are two types of excitation functions--voiced and unvoiced--and a few sounds use both together. To generate the voiced excitation function, the volume air flow through the glottis, or the post-glottal pressure, is measured by measuring glottal tissue locations using EM waves. Air flow through the area of the glottal opening can be measured during voiced speech by using EM sensors to measure the change in reflection level of the glottal region as the vocal folds open and close, and then using calibrations and models to obtain the air flow. Similarly, pressure can be measured. EM sensors measure reflection changes from the front or sides of the speaker's voice box (Adam's apple). An analytic calculation of the area opening is derived from a model functional dependence of EM reflectivity from the opening. A second technique to obtain the area is to correlate the reflected EM signal with measured optical images of the area of the opening of a representative set of speakers' glottises. A third technique is to use one or more range gated EM sensors to accurately follow the reflection from one or both edges of the glottal opening, in the sensors' line of sight, and to calibrate such signals with optical images. A fourth method is to construct a table of EM signals versus calibrated, in situ, air flow or pressure sensor signals on representative speakers during a training period.

Known equations or calibrations defining the volume air flow through the glottal opening (between the vocal folds), under conditions of constant transglottal pressure, can be used to define volume air flow vs. time in an absolute or relative fashion. This volume air flow function provides a new and valuable description of the human vocal tract voiced excitation function for each time frame of voiced speech. Similarly, post glottal air pressure can be calibrated and obtained, as needed, for correction of transglottal pressure estimates and other applications.

The change in the air flow as a function of time for the voiced excitation function can be estimated in cases when the

transglottal pressure is not constant during the time frame of estimation. This process makes use of calculated back pressure from the estimated transfer function, which is then used to make a first order air flow correction. The estimation uses models of the allowed glottal motion to determine valid glottal motions due to changes in back pressure as a function of frequency. or it uses direct measurement of tissue motions due to the pressure variations.

Acoustically generated noise can be removed from the glottal signal by using microphone information to subtract the noise signal, or by using Fourier transform techniques to filter out acoustic signals from the glottal motion signals.

The functional shape of the volume air flow excitation function in real time, and in transform space (Fourier or Z transform), can be approximated, including the glottal zero (or closed) time. An excitation feature vector is constructed by defining an approximation functional (or table) to the measured excitation function and by obtaining a series of numerical coefficients that describe the functional fitting to the numerical data for the defined time frame(s).

The number of speech frame time intervals during which both the excitation function and the acoustic output remain constant is determined. Constant is defined as the signal remaining within a band of acceptable change in real time or transform space. A feature vector can be defined describing both the excitation function and the defined number of time frames during which the two functions remain constant.

A slowly changing functional form (such as pitch period) of the volume air flow excitation function, and corresponding acoustic output, over several speech time frame intervals can also be determined, and a feature vector defined describing the excitation function and the functional changes for the defined time frames. Other slow changes such as amplitude can be similarly described.

The measured excitation function, including noise and back pressure terms, can be compared to an average speaker and a feature vector defined based upon deviations (i.e., differences) from the voiced excitation function of an average speaker or of a specific speaker.

This can be done in real time or Fourier space. Similarly, difference feature vectors can be formed by comparing a recently obtained featured vector to one obtained from an earlier time frame.

The invention also includes using the voiced excitation function periods as master timing units for the definition of time frames during speech processing. This includes defining the beginning and end of a glottal open-close cycle, obtaining the times of glottal closure (i.e., no air flow) within the cycle, and joining one such cycle to the next for concatenation of all information obtained in one speech time frame to that obtained in the previous or next time frame.

Single or multiple time frame timing unit measurements can be made of simultaneous speech organ conditions and other conditions such as video, electrical skin potential, air flow, magnetic resonance images, or ultrasonic wave propagation.

The invention includes characterizing and storing as part of a feature vector the automatically generated time frame information; associating each speech time frame with a continuous timing clock, and storing this absolute timing information as part of a feature vector; and using such defined time frames for the purposes of speech reconstruction, speech synchronization with visual images, visualization of vocal organ conditions for training or speech prosthesis, speaker identification, foreign language translation, and coded telephony.

The invention includes methods to estimate the unvoiced excitation functions of the speaker during defined speech time frames, by determining that speech is occurring without vocal fold motion. A "modified white noise" excitation function is then selected from a functional form that has been validated by listeners and by analysis to provide an accurate excitation function to excite the known transfer functions of average speakers (in the language of the speaker) to simulate the measured acoustic output for known sounds. A second method is to deconvolute the known transfer function for the unvoiced sound from the acoustic output and obtain a measured unvoiced excitation function source.

Speech unit time frames are defined when unvoiced speech is being sounded by the speaker during the speech time frames of interest. The algorithm is to simply measure the time duration over which the acoustic spectrum is constant and record that time to be the frame duration; or, using spectral constancy, and times defined by extrapolated or interpolated voiced-speech time frame duration from the preceding or following voiced speech periods; or by using pre-defined time frame periods, e.g. 50 ms.

A preferred unvoiced-excitation-function feature-vector is defined by the Fourier transform for one or more speech time frame intervals during which the excitation function is constant or slowly varying. The number of unvoiced speech frames during which a constant or slowly changing unvoiced excitation of the vocal tract is occurring is determined, and a feature vector is defined that describes the excitation function, the time frame duration, and the slow changes in the excitation function over the defined time frames.

The invention includes a method of measuring and recording the acoustic output of the human speaker, simultaneously with the EM sensor signals, during one or more speech time frames and storing the information with sufficient linearity, dynamic range, and sampling bandwidth for the user's application .

The microphone voltage amplitude vs. time signal recorded during the speech time interval frame or frames is characterized in real time or in Fourier frequency space for the purpose of deconvoluting the excitation function from the recorded acoustic output function. Information is selected from the recorded microphone voltage vs. time signal that is statistically valid and characterizes the sound pressure amplitude vs. time or the sound pressure Fourier amplitude and phase vs. frequency during the desired time frame (s) for the purposes of subsequent processing. The lip-to-microphone acoustic radiation transfer function can be deconvoluted, in Fourier space or in real time space, to remove instrument artifacts, to simplify the transfer function, and to enable more rapid convergence of deconvolution procedures in subsequent processing steps.

The invention includes a method of using EM speech organ position or velocity information (e.g., vocal folds) for one or several sequential speech time frames to deconvolve the vocal system source function from the measured acoustic speech output from a human speaker. This makes possible an accurate numerical representation of the transfer function of the human vocal tract in use during the time frame(s) over which deconvolution is performed. Deconvolving can be done by real time, by time series techniques, by fast Fourier transform techniques, by model based transform techniques, and other techniques well known to experts in the field of data processing and deconvolution.

A human speaker's vocal tract transfer function used during one or more speech time interval frames is obtained by using well known deconvolution techniques (such as that associated with the ARMA approach) by dividing the transformed microphone acoustic pressure signal by the transformed excitation source signal. The lip to microphone transfer function, or other known functionals, can be obtained as needed by deconvolving, fitting to known functionals, or other well known numerical techniques.

Additional information on the positions of individual organ locations, and thus the shape of the vocal tract, can be obtained through the use of other EM sensor data, with or without simultaneous acoustic data, to determine the optimal transfer function functional structure for best convergence or most accurate fitting. An example is to choose the appropriate number of poles and zeros in the ARMA functional description for each speech time interval frame.

A speech transfer-function feature-vector can be defined from the amplitude and phase vs. frequency intervals from the deconvolving of the excitation function from the acoustic output function, using Fourier transform or other techniques. The function can be defined by a table of numerical values or be fit by a known functional form and associated numerical parameter coefficients.

The invention includes a method of approximating the transfer function by using the well known pole-zero (or time series a, b coefficient) approximation techniques such as used by the auto

regressive-moving average (ARMA) technique. Transfer function feature vectors are formed for the speech time interval frame or frames, including obtaining amplitude, phase, type of functional form, defining functional coefficients, time duration of feature vector, and other necessary information.

A feature vector describing the transfer function is formed by using the pole and zero representation or the a, b representation of the ARMA description for the speech time interval frame or frames of interest. A feature vector describing the transfer function is also formed by using defined ARMA functional forms which are based upon fixing the numbers of poles and zeros to be used (or alternatively the a, b values) of the ARMA description for the speech time interval frame or frames of interest.

The invention includes defining a difference "Pole-Zero" (or a, b) feature vector by storing differences in each vector element from a previously defined known type of speaker or by storing differences from past time frames during a constant period of use. It also includes the definition of "more important" pole-zero (or a,b) values which define major tract dimensions, and "less important" values which define the idiosyncratic sounds of an individual human speaker.

The invention includes approximating the transfer function by using well known electrical and/or mechanical analogies of the acoustic system which are predefined by foreknowledge of the human vocal tract acoustic system, including transfer function "feature-vector" formation for the speech time interval frame(s). Feature vectors describing the transfer function are formed by using the impedances, (i.e., the Z's), or circuit values (e.g. L's, C's, R's, G's) in the electrical analog models. A feature vector can be defined by storing differences in each vector element from a previously defined known type of speaker, or from coefficients obtained in a previous time frame.

The feature vector and excitation function information can be used to define the physiological parameters of the human speaker. The transfer function parameters are used to define the electrical analog models and are associated with physiological parameters such as tract length, mouth cavity length, sinus volume, mouth volume, pharynx

dimensions, air passage wall compliance, and other parameters well known to acoustic speech experts. The excitation function information can be used to define the masses, spring constants, and damping of the glottal membranes.

A feature vector describing the transfer function can be formed by using the physiological dimensions of the speaker that are defined by the measured and derived transfer functions for the vocal tract configurations and used by the speaker during the speech time interval frame or frames of interest. A feature vector is also formed by storing differences in each feature vector element from a previously defined known type of speaker as a feature vector, or from coefficients taken in a previous time frame.

The invention includes a method of defining for each time frame and for multiple time frames, a sound feature vector that is a "vector of vectors". It is comprised of the user defined needed information from the excitation function feature vectors, vocal tract transfer function feature vectors, prosody feature vectors, acoustic feature vectors, timing information, and control information for all acoustic sound units, over as many time frames as needed, for the application in the language of use. It includes obtaining and storing such vectors in a data base (i.e. library or code book) during training sessions. The data bases are designed for rapid search and retrieval during real time usage. This method includes defining each unique speaker, defining reference speakers using individuals or averaged speaker groups, or translating coefficients to a hypothetical speaker using normalization, or artificial modifications of the functionals and their coefficients. It also includes forming such a vector over one or more defined speech frames, which includes the formation of the above for all syllables, phonemes, PLUs, diphones, triphones, multiphones, words, phrases, and other structures as needed in the language of use and for the application.

The stored feature vector information, contained in the type of functional and the defining feature vector coefficients on a given speaker can be used to normalize the output of the subject speaker to that of an average speaker. This normalization method recognizes the

differences of an individual by comparing his individual excitation function and transfer function coefficients for known sounds, to those of a reference speaker's excitation function and transfer function coefficients, for the same sound during training sessions. The simplest method is the method of replacement of reference speaker feature vectors with those of the user and a second method is to replace feature vectors describing difficult sound combination. These personalize the code books and make comparison more accurate, and retrieval of vectors very individualized. A third method is a method of extremes, in which a mapping is made from the extremal values of each coefficient in the feature vector of the user to those a reference speaker. The values include the coefficient range-extremes for all necessary sound units for the application, and are obtained during training. Then feature vector coefficients obtained each time frame are normalized to those of the reference speaker by using a linear fractional mapping. This approach removes much of each individual's articulation variability, and allows the formation of a speaker independent feature vector for each time frame. In this manner, a speech sound can be associated with a sound symbol in a stored library with very low ambiguity and very high probability of identification. This approach also removes instrument variations.

The method includes quantizing the normalized feature vector coefficients into a limited set of values that reflect bands-of-distinguishability for the application. It is known that articulators must change their position or condition a certain amount for a noticeable speech difference to be considered important by the user. The bands of coefficient values that are perceived to be constant, are measured during system set-up and during training. As each normalized coefficient is obtained, it is mapped into one of a few values that reflect the "quantized" aspects of the speech articulator. This approach makes possible very rapid table look up, using the coefficients themselves to directly access codebook addresses for the corresponding stored reference feature vector .

The complete feature vector for several time frames, over which slow change or no change at all in the vector coefficients, can be

collapsed to a feature vector describing one speech frame. In addition, the collapsed feature vector contains a few additional coefficients describing the total recorded duration of the sequence of constant time frames, plus some that define a model of the slow changes in one or a few coefficients over the entire sequence. This procedure is a method of speech compression that removes redundant information, and yet retains as many of the speaker's qualities as desired for the application.

The complete feature vectors, for one or more time frames, can be compared to stored information on a known human for the purpose of speaker identification, and providing statistics of identification. Such comparisons can be performed automatically over several time frame units, isolated time frame units, or on sequences of units where stored information on the desired speaker's identity is available from a preformed library. The speaker can speak prearranged words or can respond to information presented by the system, or the system can recognize sequences of units, using speech recognition, and compare them to stored information on the desired speaker's identity obtained from a pre-formed library.

The invention provides a method to code an individual's speech, not knowing the language being spoken, and to search through a series of code books for one or more languages to identify the language being spoken. The process makes use of the statistics of each language's sounds, sound patterns, and special unique sounds to obtain the language recognition.

The invention includes a method of speech recognition based upon using the feature vectors for the purposes of identifying all sound units in a given language. The simplest recognition technique, directly applicable with the methods herein because of their accuracy, is often called a phonetic template approach. A feature vector describes the condition of a speech unit with sufficient information, including redundancy and model constraints, that the phoneme (or other simple speech sound unit) of speech can be defined for the time period and be directly matched to a pre-formed vector stored in a codebook.

The sound unit under consideration, once identified with very high probability, is associated with a symbol. Symbols can be letters,

ASCII computer code, pictogram symbols, telephony code, or other coding known to practitioners of speech recognition, synthesis, telephony and similar activities.

The invention includes a second method of speech recognition that uses Hidden Markov Model (HMM) techniques on a multi-time-frame feature-vector to statistically identify the sequence of phonemes being spoken in the examined time frames. The feature vectors are so accurate that this approach becomes fast, accurate, and accommodates large natural language, continuous speech vocabularies. This includes a learning phase as is well known for the HMM approach to conventional speech recognition. HMM techniques can be used to identify the diphones, triphones, multiphones, words, and word sequences in the examined time frame.

The invention includes a method of using joint probability on the feature vectors to statistically identify the phoneme being spoken in the examined time frame using multiple sensor input. Joint probability includes the use of a conventional speech recognition technique for the first step. It estimates the identify of one or more sound units and it records its probabilities of identification for the next step. The second step is to use the EM/acoustic defined feature vectors, obtained by deconvolving, to estimate separately the identity of the sound unit, and to assign a second set of probability estimates for the nonacoustic case. A third step uses EM sensor information alone and a third set of identified speech units and their probabilities are formed. The final step is to join the probabilities of each estimate to obtain a more accurate identification of the word unit than either an all acoustic system, an EM/acoustic, or an all EM feature vector system could accomplish by themselves. The joint probability technique can identify the diphones, triphones, multiphones, words, and word sequences in the examined time frame.

The invention also includes a method of using exclusive probability on the feature vectors to statistically differentiate between acoustically similar phonemes being spoken in the examined time frame using several different sensor information sets. Exclusive probability means starting, for example, with a conventional speech

recognition technique to estimate the identity of one or more sound units. They may have similar probabilities of being defined using conventional acoustic techniques alone (i.e. there remains ambiguity in a statistical sense). The second step is to use, for example, the EM/acoustic defined feature vectors of each of the one or more acoustically identified phonemes to estimate separately the identity of the sound units, and to assign an estimate of the probability based on EM/acoustic generated vectors for each ambiguous sound unit. Any sound unit from the first step that does not meet a minimum probability from the second step, is removed from further consideration (i.e., it is excluded). This reduces computational time, because those units that are rejected early, are no longer considered. A third step can use EM sensor information alone, to test the remaining sound units from steps 1 and 2, and if they do not meet the criteria, they are rejected. A final step is to join the probabilities of each estimate to obtain the most accurate identification of the remaining word unit or units, than either an all acoustic system, or an all EM/acoustic feature vector system could accomplish. In this manner, one can exclude all of the units identified from the first step (e.g., acoustically identified sound units in this example) except for one that meets the criteria defined by comparison with the library of stored feature vectors for the following steps. The order of sensor approach can be interchanged. The exclusive probability technique can identify the diphones, triphones, multiphones, words, and word sequences in the examined time frame.

The invention includes a method of using neural network algorithms to associate a pattern described with the feature vectors in conjunction with the symbolic representation of the corresponding sound units. This method uses the usual training methods for neural networks (including normalization and quantization of input feature vectors), the averaging of speakers (one or more), and associating the inputs though the neural network algorithms (back propagation, two or more layers, etc.) with known words or other speech units. Once trained, the networks provide a rapid association of an input feature vector to an identified output speech unit symbol because the input data

from the methods are so well defined, speaker independent, and accurate.

The invention includes a method of synthesizing high quality, idiosyncratic speech from stored EM sensor obtained data for an individual speaker. Individual speaker means coding the speech of an average office dictation worker or a famous actor. The quality of the speech depends upon the quality of the coding of the original feature vectors, their storage in a code book, and the retrieval methods and concatenation methods. First the needed speech units are recorded, coded, and stored with associated symbols in a code book. Second, a commercial text to speech translator is used that identifies all of the required speech units (phonemes, diphones, triphones, etc.) from written text for the purpose of retrieving the desired speech feature vectors from the code book. Next the sound units to be used, the timing of the units, and the prosody are selected. The units are joined together by convoluting the excitation functions with the transfer functions to produce the output sound function, and using, in the preferred embodiment, the period of glottal closure as the timing "mark" for joining speech interval segments. Finally prosody is provided for each speech unit or combination of speech units; in particular it sets the sound level, and the pitch change from the beginning of the unit to the end as defined by phrasing and punctuation. Other concatenation approaches can be used as well, because the procedures allow easy selection of function values and derivatives.

The invention includes a method of altering the synthesized speech by altering the stored speech feature vectors. The pitch is changed by modifying the excitation function feature vector by increasing the number of glottal open and close cycles per unit time, and then convoluting this higher pitch excitation with the vocal tract transfer functions for each defined length feature time interval. This is done by compressing the descriptors of the excitation function so that a similar, but shortened pattern, in time, is derived. The individual speech feature vector can be altered to a predefined normalized speech vector. In addition, speech duration can be shortened or lengthened by

adding or subtracting speech frames, including silence periods, in units of glottal periods.

The transfer function of the speaker can be altered in a known way by altering the physiological parameters in a known way, such as lengthening the vocal tract or increasing the size of the nasal cavity based upon the automatically derived data. Once the physiological parameters are changed, then a new transfer function feature vector (along with excitation and prosody vector elements) is formed based upon the new physiology of the vocal tract for the time frame being investigated.

The excitation function of a more desirable speaker, or the transfer function, or the prosody pattern for a given speaker can be substituted, before performing the convolution, upon demand, for the purpose of improved speech synthesis.

Synthetic excitation functions (e.g. unphysical open-close shapes, or very high pitch) can be generated, or non-physical modified transfer functions (e.g. based upon exaggerated physiological parameters) or amusing or desirable prosody patterns for the purposes of entertainment, speech research, animal research or training, or specially desired effects.

The invention includes using these coding techniques for the purposes of coding the feature vectors of a speaker speaking into a telephony set transmitter microphone. This coding includes attaching additional information as desired such as speaker identification, speech alteration if needed, and translating the feature vectors into appropriate code for transmission. The real time speech recognition of the speech can occur and the corresponding symbol can be identified, and transmitted with dramatic drop in bandwidth. These methods allow simplified encryption, foreign language translation, and minimal bandwidth coding for the transmission of the coded units via wire, optical fiber, or wireless in real time. The methods include how to synthesize the coded speech (e.g., symbols or feature vectors) into acoustic speech representing the speaker for broadcasting the rendered acoustic sounds through the telephony receiver to the listener. The speech synthesis can also be designed to use for identifying, sending,

and/or synthesizing prestored average speaker qualities, to send "difference feature vectors", to send partial information using "most important" and "less important" functional fitting terms. It can be designed to transmit very high fidelity speaker idiosyncratic speech, and thereby use relatively higher bandwidth for the transmission of the more accurate description of the feature vector information, or minimal quality to minimize bandwidth.

The inverse communication channel works in the same fashion, except the listener becomes the speaker and the speaker the listener. Real time means that the recognizing, coding, and synthesizing can take place while speakers are speaking or while speech is being synthesized and with a time delay that is short enough for the users to be satisfied.

The invention also includes telephone coding using identification procedures where the speech recognition results in a word identification. The word character computer code (e.g. ASCII) is transmitted along with none or minimal speaker voice characterization information for the purpose of minimizing the bandwidth of transmission. Word (i.e., language symbols such as letters, pictograms, and other symbols) transmission is known to be about 100 fold less demanding of transmission bandwidth than present speech telephony; thus the value of this transmission is very high.

The methods include communication feedback to a user for many applications because the physiological as well as acoustic information is accurately coded and available for display or feedback. For speech correction or for foreign language learning, displays of the vocal organs show organ mispositioning by the speaker. For deaf speakers, mis-articulated sounds are identified and fed back using visual, tactile, or electrical stimulus units.

Changes and modifications in the specifically described embodiments can be carried out without departing from the scope of the invention which is intended to be limited only by the scope of the appended claims.

THE INVENTION CLAIMED IS

1. A method for characterizing speech, comprising:

directing EM radiation toward speech organs of a speaker;

detecting EM radiation scattered from the speech organs to obtain speech organ information;

detecting acoustic speech output from the speaker to obtain acoustic speech information;

combining the EM speech organ information with the acoustic speech information using a speech coding algorithm to obtain the speaker's excitation function and speech tract transfer function.

2. The method of Claim 1 further comprising defining a speech time frame.

3. The method of Claim 2 further comprising defining the time of start, stop, and duration of the speech time frame.

4. The method of Claim 2 further comprising forming feature vectors for each speech time frame.

5. The method of Claim 1 further comprising deconvolving the speech excitation function from the acoustic speech information to produce a deconvolved transfer function.

6. The method of Claim 5 further comprising forming a feature vector by fitting the deconvolved transfer function to a mathematical model.

7. The method of Claim 6 wherein the feature vector is formed by one of numerical table look-up, Fourier transform, an ARMA model technique, an electrical or mechanical analog model of the acoustic system, or an organ-dimension physiological/acoustic-model of the acoustic system.

8. The method of Claim 6 further comprising choosing the transfer function mathematical model using EM sensor information describing the dimensions and locations of vocal organs.

9. The method of Claim 8 further comprising obtaining the transfer function using real time measurements.

10. The method of Claim 1 wherein the EM radiation is directed to and reflected from the glottal region and is sensed in the near field mode, the intermediate field mode, or the far field mode.

11. The method of Claim 2 wherein the speech time frame is defined by measuring glottal opening and closing using reflected EM waves.

12. The method of Claim 11 further comprising defining a composite time frame from two or more glottal opening and closing time frames.

13. The method of Claim 11 further comprising precalibrating an EM sensor so that the EM signals can be converted to either pressure and/or volume air flow in real time.

14. The method of Claim 11 wherein a voiced excitation function feature vector is described by numerical table values or by fitting a mathematical functional model to the numerical table values.

15. The method of Claim 2 comprising obtaining the excitation function for unvoiced speech.

16. The method of Claim 15 comprising defining an unvoiced speech time frame by the absence of EM detected glottal opening/closing and the presence of acoustic output.

17. The method of Claim 11 comprising forming the feature vector for combined voiced and unvoiced speech time frames.

18. The method of Claim 4 further comprising forming difference feature vectors.

19. The method of Claim 6 further comprising dividing the transfer function into "important" pole-zero terms describing major vocal tract configurations and "less-important" pole-zero terms describing idiosyncratic speaker's vocal organ physical and acoustical conditions.

20. The method of Claim 4 further comprising comparing a feature vector to stored feature vector information to identify a speaker.

21. The method of Claim 4 further comprising comparing a feature vector to stored feature vector information in many language

codebooks to identify the language being used by the speaker for the formation of acoustic speech units.

22. The method of Claim 4 further comprising normalizing the feature vector of a speaker to that of one or more reference speakers.

23. The method of Claim 4 further comprising quantizing a continuous coefficient-value band of a feature vector to a small number of distinct coefficient values representing a small number of distinct user-discernible, application-related speech conditions defined by each coefficient.

24. The method of Claim 4 further comprising defining acoustic speech unit feature vectors by combining one or more excitation function feature vectors, vocal tract transfer function feature vectors, prosody feature vectors, timing, algorithm control coefficients, neighboring frame connectivity coefficients, and acoustic feature vectors for all acoustic units in a language.

25. The method of Claim 24 further comprising generating said combined feature vectors with identifying symbols for all acoustic speech units used in a language and storing them in a library, codebook or data base.

26. The method of Claim 24 further comprising averaging feature vector coefficients from the excitation, transfer, acoustic, prosody, and timing functions of one or more speakers to form a reference speaker acoustic sound unit feature vector and storing them in a codebook or data base.

27. The method of Claim 24 further comprising modifying feature vector coefficients and functional representations of the excitation, transfer, acoustic, prosody, neighboring frame connectivity, and timing functions of one or more speakers to form a modified acoustic sound unit feature vector and storing them in a codebook or data base.

28. The method of Claim 25 further comprising associating a foreign language word or phrase symbol in a second language with each unit of a first language coded by a speaker or speakers and storing them in a codebook or data base.

29. The method of Claim 24 further comprising storing the acoustic speech unit feature vectors in a library, code book, or database.

30. The method of Claim 4 further comprising identifying all sound units in a language from the feature vectors.

31. The method of Claim 30 further comprising identifying all acoustic speech units in a language by a method selected from the group consisting of template matching techniques, HMM techniques, neural network techniques, a method of joint probabilities of two or more identifying algorithms, and a method of exclusion to reject identified units in a sequence of tests by two or more identifying algorithms.

32. The method of Claim 30 further comprising identifying each acoustic speech unit with a symbol of the language unit identified.

33. The method of Claim 1 further comprising synthesizing speech from the EM and acoustic speech organ information.

34. The method of Claim 33 wherein speech is synthesized by:

generating a code book of reference speaker feature vectors and identifying symbols;

identifying speech units for synthesis using a text to speech translator;

selecting the sound units and timing;

providing selected sound feature vectors from a stored data base;

concatenating the sound units in speech sound sequences;

modifying feature vector coefficients or sequences of feature vector coefficients using prosody rules;

modifying the time duration of individual sounds; and

generating sound feature vectors by convolving the modified excitation functions with the modified transfer functions to produce an output sound function.

35. The method of Claim 34 further comprising measuring positions on an excitation function amplitude versus time function to join speech interval segments together.

36. The method of Claim 35 further comprising using a time during glottal closure as a timing marker for joining speech frame segments.

37. The method of Claim 1 further comprising coding acoustic speech units, transmitting the codes to a receiver system, and reconstructing the transmitted codes to acoustic speech.

38. The method of Claim 37 wherein the codes are symbolic codes.

39. The method of Claim 37 further comprising modifying the codes to transmit minimal information, and reconstructing the codes to acoustic speech using locally stored code books of reference speakers.

40. The method of Claim 37 further comprising obtaining an associated foreign language symbol or speech code, transmitting the foreign language code to the receiver system, and reconstructing to acoustic speech in the foreign language.

41. The method of Claim 37 further coding the acoustic speech units in a first language, transmitting the coded information from the first language, recognizing the transmitted coded units, obtaining associated language symbols or speech codes in a second language from a system codebook at the receiver system, and reconstructing acoustic speech in the second language at the receiver system.

42. The method of Claim 4 further comprising communicating back to the speaker or to others speech organ articulation qualities, which are coded in the feature vectors for the speech time frames, by using communication vehicles selected from the group consisting of visual images, printed information, acoustic messages, and tactile and/or electrical stimulus.

43. The method of Claim 24 where a speech segment is compressed by:

forming a sequence of feature vectors for each sequential time frame in the speech segment;

comparing sequential changes in the feature vector coefficients, for each feature vector in the sequence, against a predefined model describing change in one or more of the coefficients over the sequential time frames;

forming a single representative feature vector for several time frames over which the coefficients meet the criteria of the predefined model;

adding to the representative feature vector extra coefficients describing the predefined model and a parametric fit to the model;

adding the total duration time of the several time frames to the representative, multi-time frame feature vector as an extra coefficient;

storing or transmitting the compressed segment electronically.

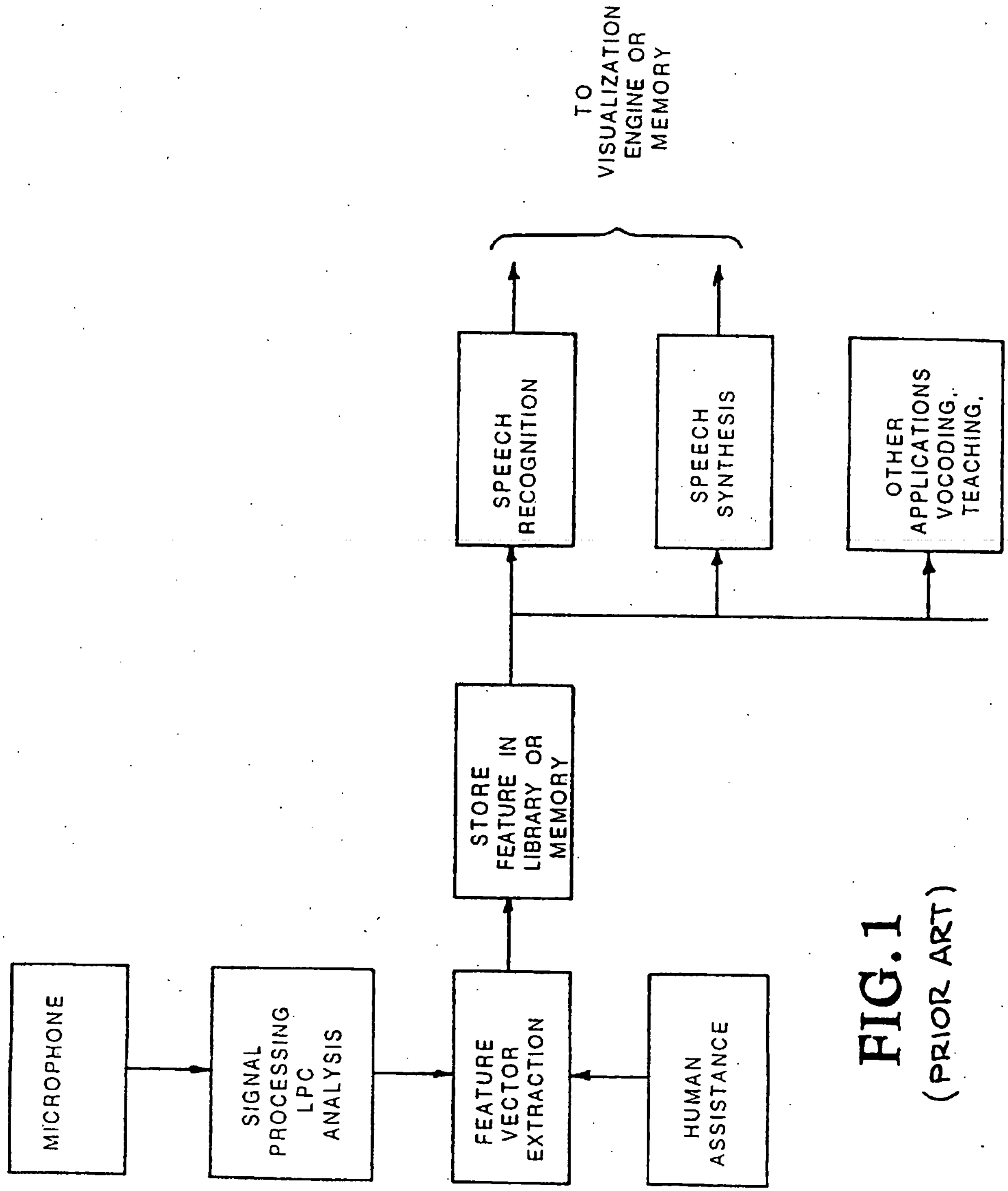


FIG.1
(PRIOR ART)

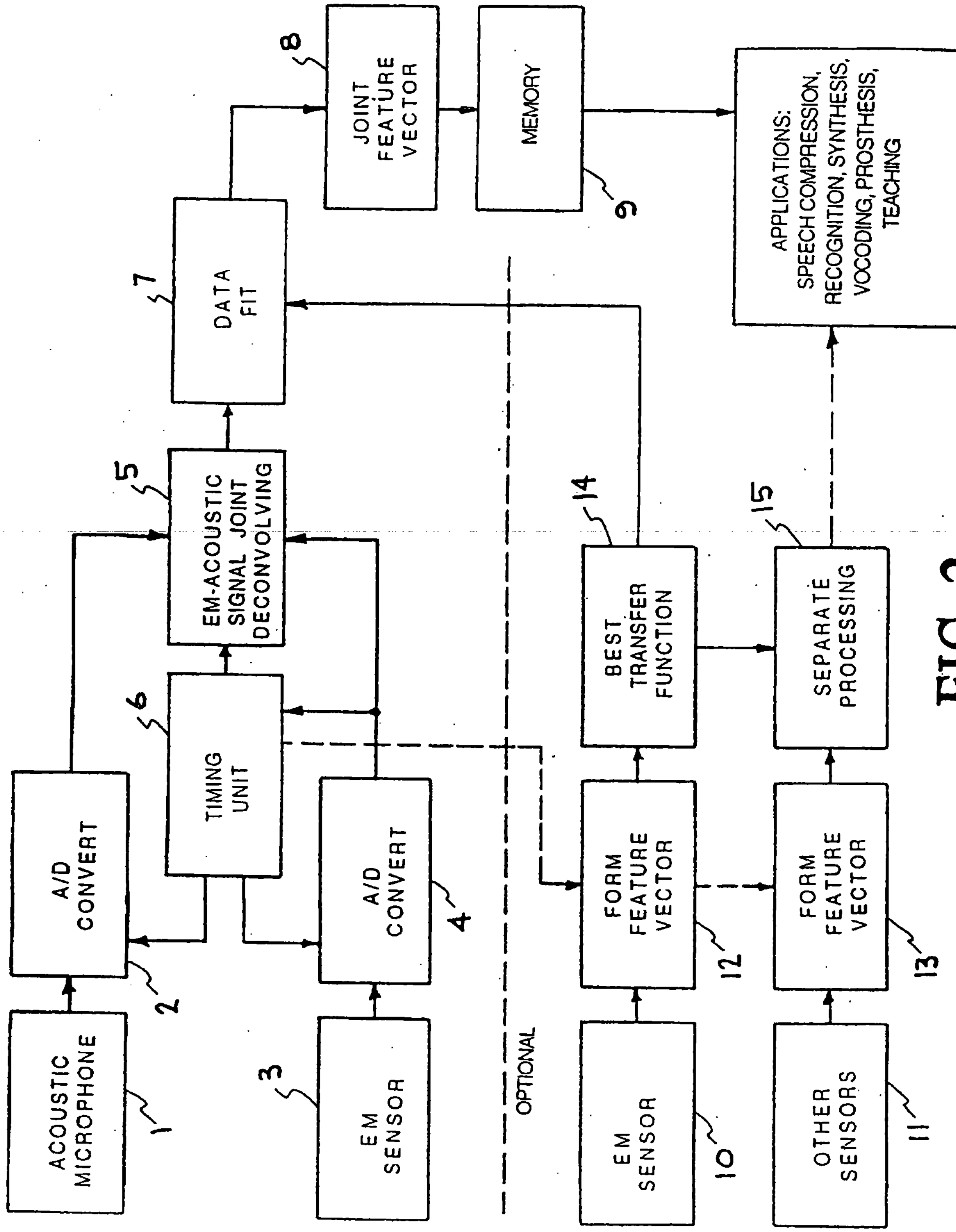

ACOUSTIC MICROPHONE
1
A/D CONVERT
2
EM SENSOR
3
A/D CONVERT
4
TIMING UNIT
6
EM-ACOUSTIC SIGNAL JOINT DECONVOLVING
5
DATA FIT
7
JOINT FEATURE VECTOR
8
MEMORY
9
OPTIONAL
EM SENSOR
10
OTHER SENSORS
11
FORM FEATURE VECTOR
12
FORM FEATURE VECTOR
13
BEST TRANSFER FUNCTION
14
SEPARATE PROCESSING
15
APPLICATIONS: SPEECH COMPRESSION, RECOGNITION, SYNTHESIS, VOCODING, PROSTHESIS, TEACHING

FIG. 2

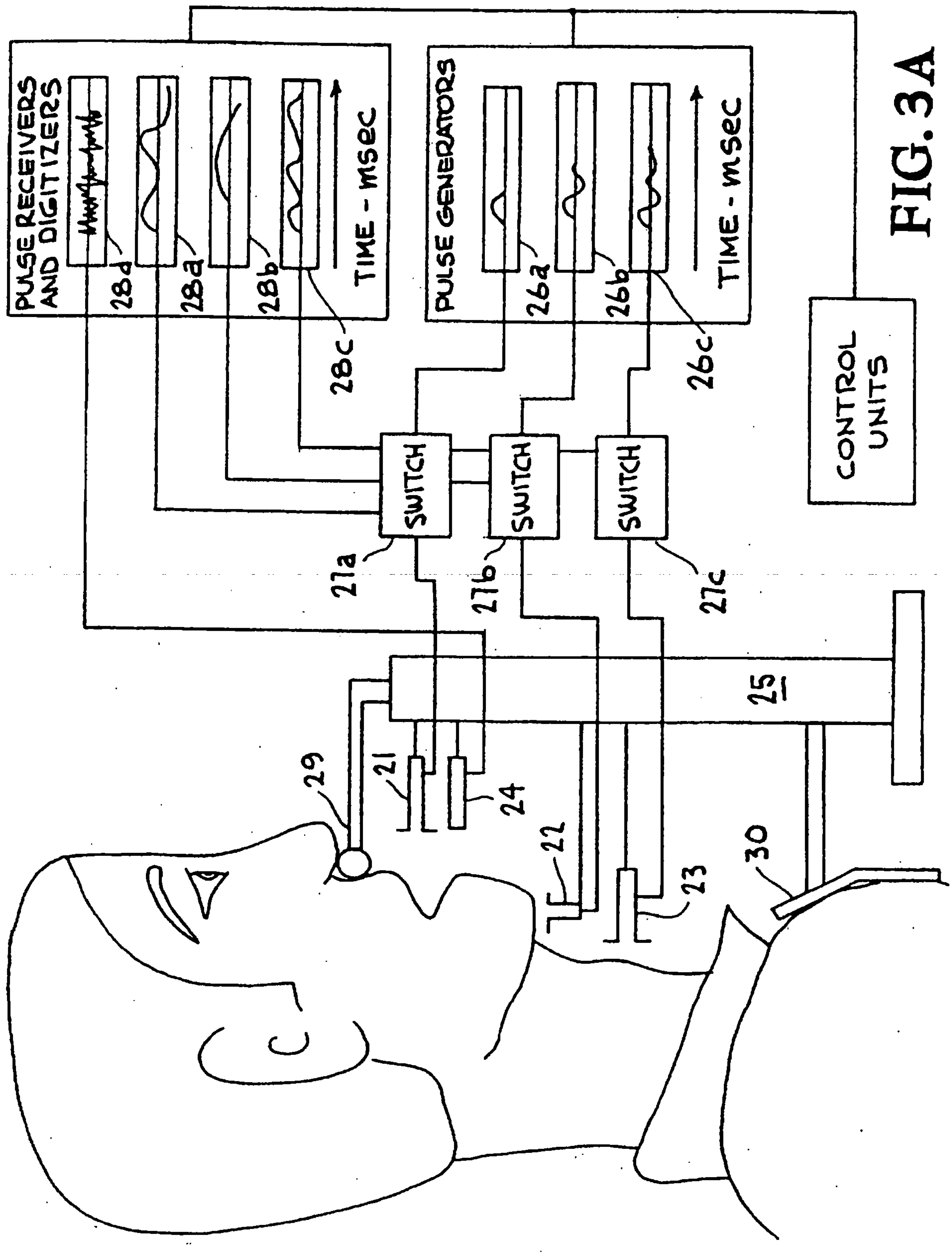

PULSE RECEIVERS AND DIGITIZERS
28d
28a
28b
28c
TIME - msec
PULSE GENERATORS
26a
26b
26c
TIME - msec
CONTROL UNITS
SWITCH
SWITCH
SWITCH
27a
27b
27c
25
29
21
24
22
23
30

FIG. 3A

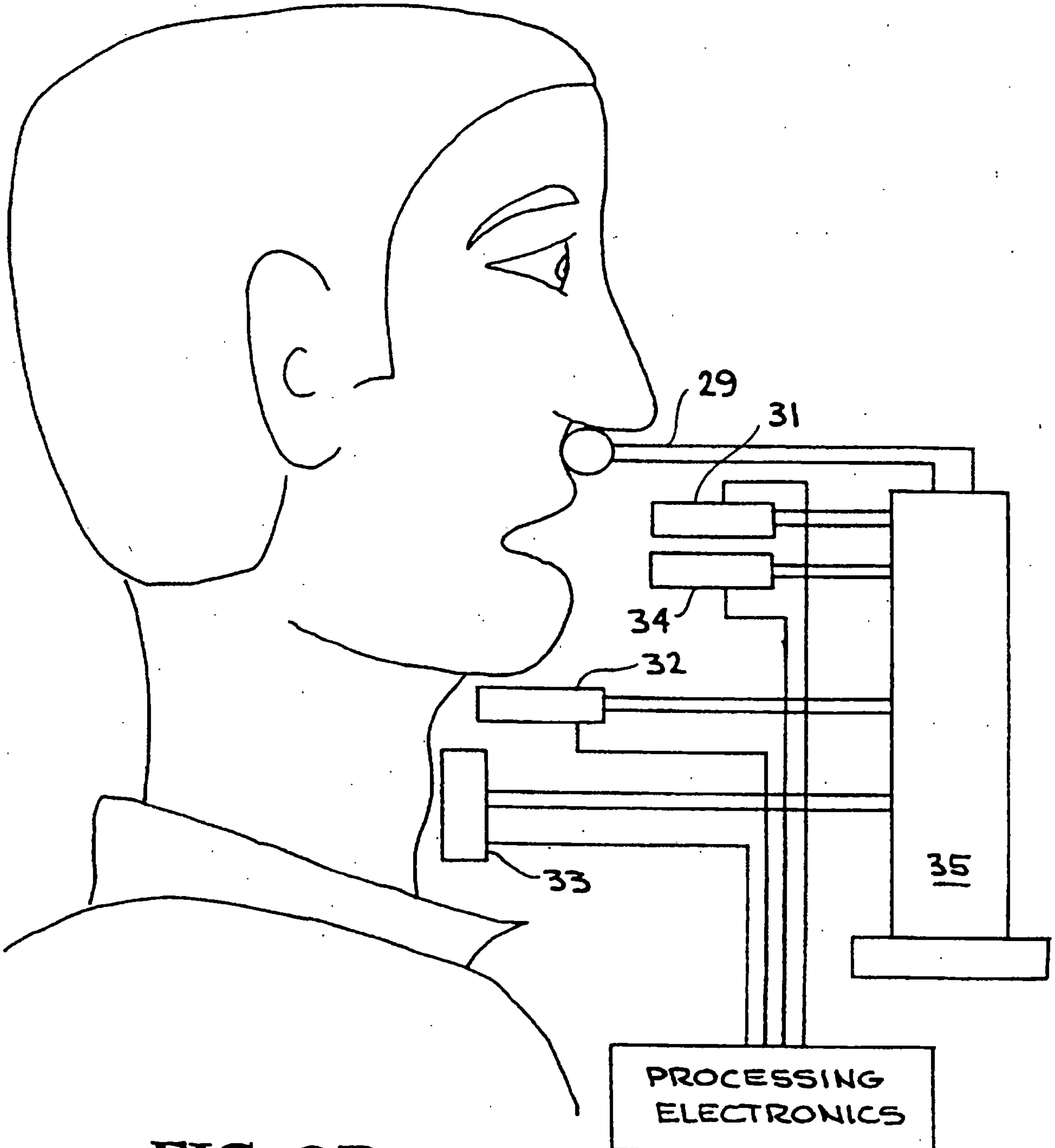


FIG. 3B

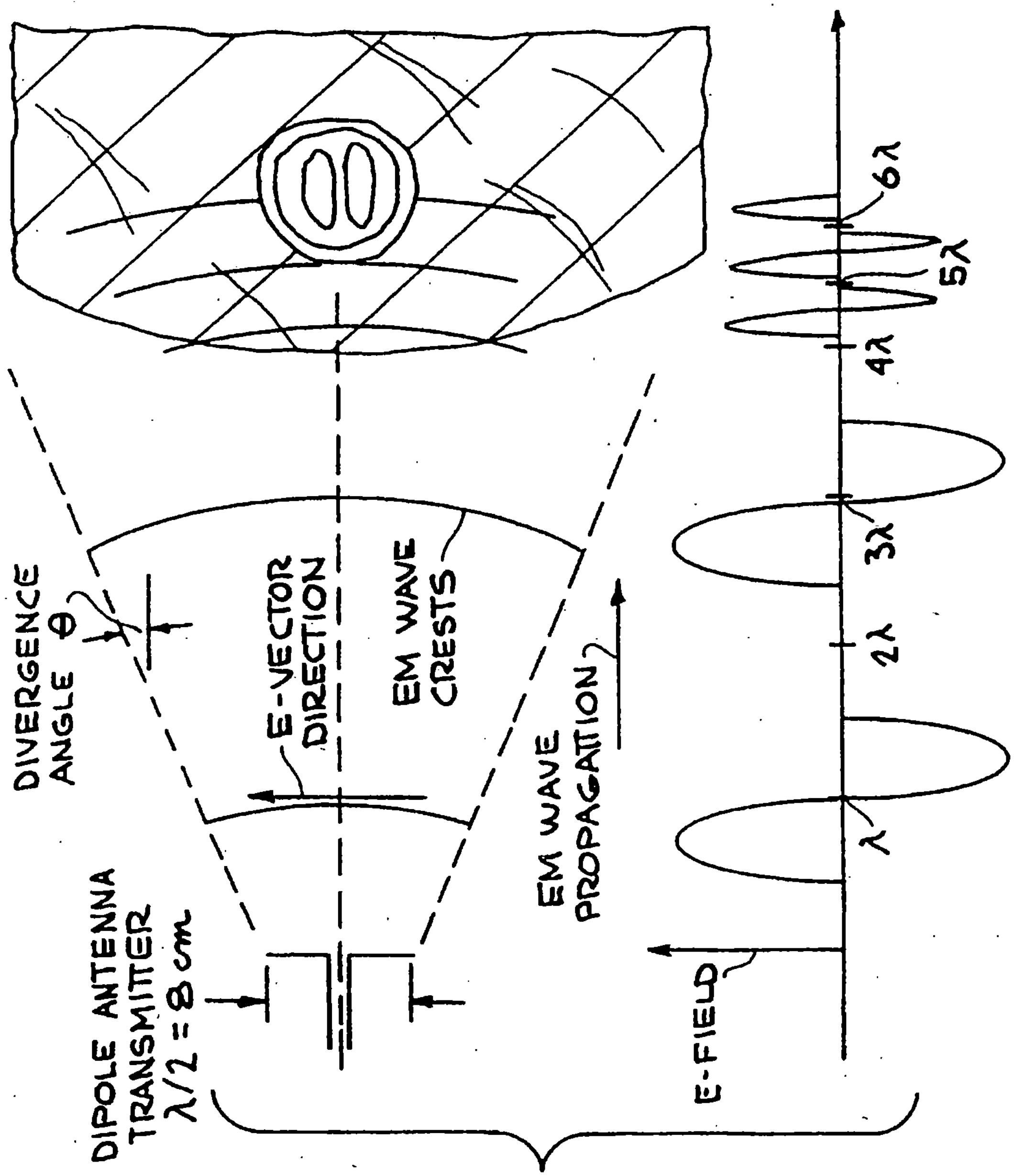

DIVERGENCE ANGLE θ
E-VECTOR DIRECTION
EM WAVE CRESTS
EM WAVE PROPAGATION
DIPOLE ANTENNA TRANSMITTER
λ/2 = 8 cm
EM WAVE PROPAGATION
E-FIELD
λ
2λ
3λ
4λ
5λ
6λ

FIG. 4

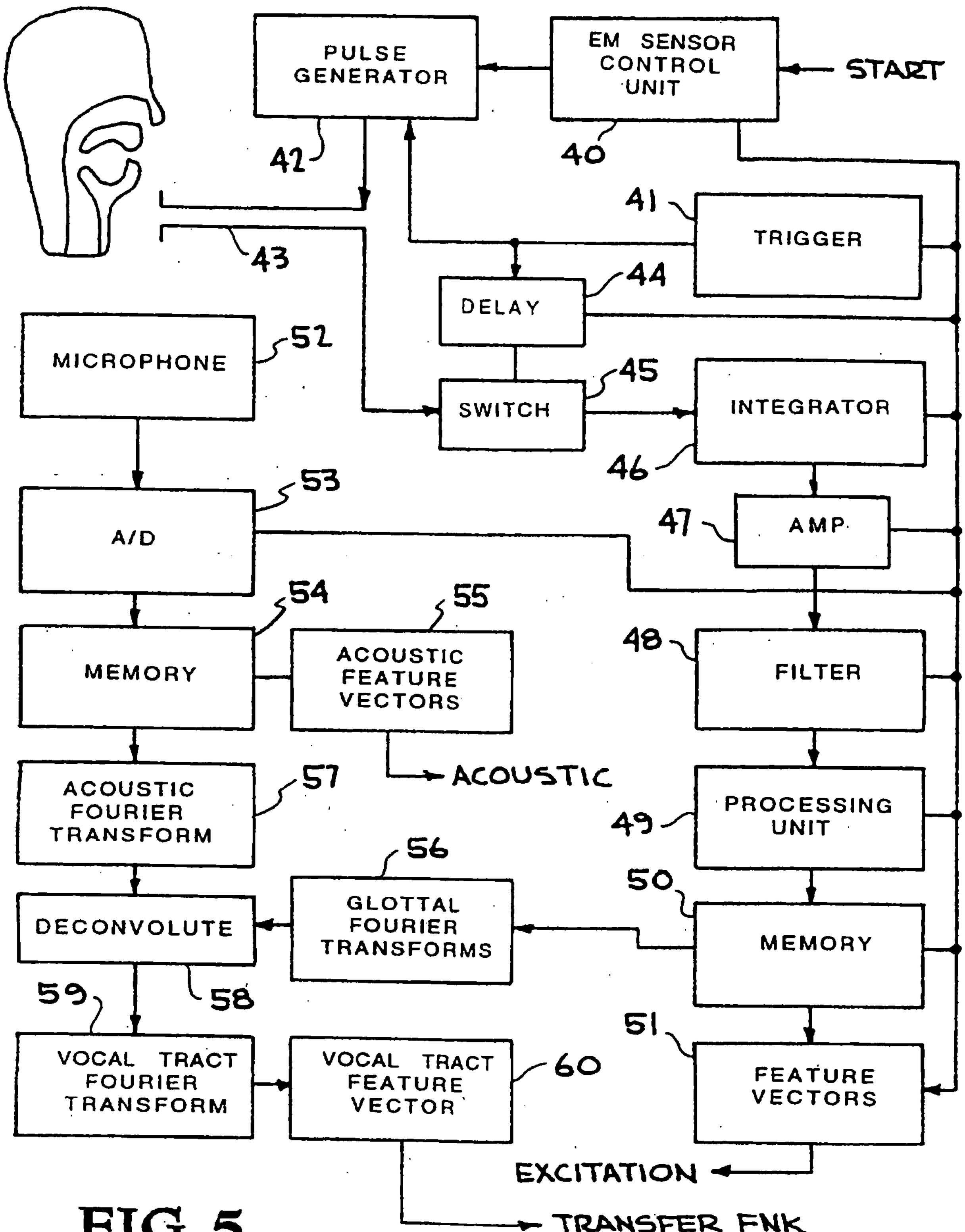

PULSE GENERATOR
EM SENSOR CONTROL UNIT
START
42
40
41
TRIGGER
43
44
DELAY
MICROPHONE
52
SWITCH
45
INTEGRATOR
46
53
A/D
47
AMP
54
55
MEMORY
ACOUSTIC FEATURE VECTORS
48
FILTER
ACOUSTIC FOURIER TRANSFORM
57
ACOUSTIC
PROCESSING UNIT
49
56
50
DECONVOLUTE
GLOTTAL FOURIER TRANSFORMS
MEMORY
59
58
51
VOCAL TRACT FOURIER TRANSFORM
VOCAL TRACT FEATURE VECTOR
60
FEATURE VECTORS
EXCITATION
TRANSFER FNK

FIG.5

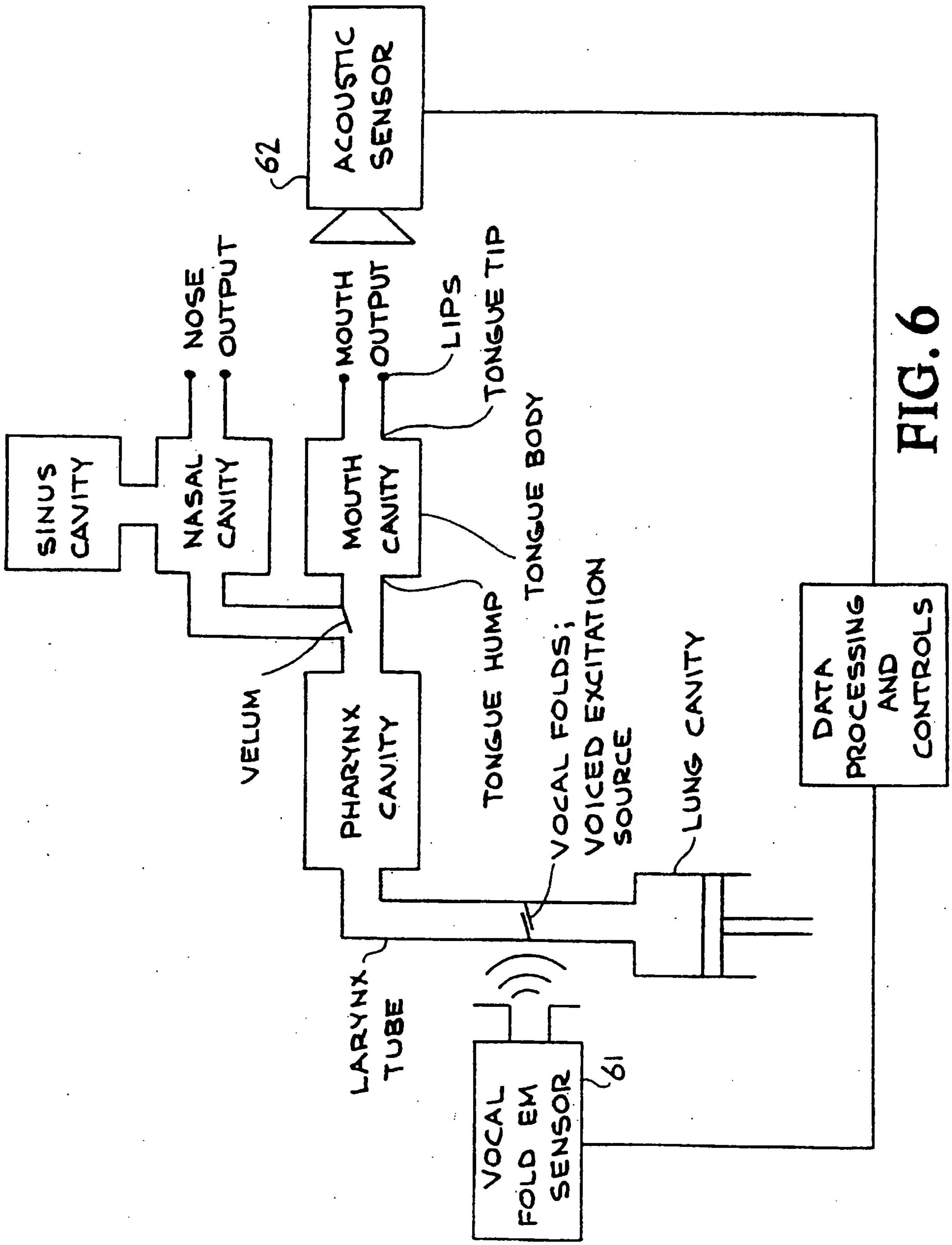

ACOUSTIC SENSOR
62
NOSE OUTPUT
MOUTH OUTPUT
LIPS
TONGUE TIP
SINUS CAVITY
NASAL CAVITY
MOUTH CAVITY
TONGUE BODY
TONGUE HUMP
VELUM
PHARYNX CAVITY
VOCAL FOLDS; VOICED EXCITATION SOURCE
LUNG CAVITY
DATA PROCESSING AND CONTROLS
LARYNX TUBE
VOCAL FOLD EM SENSOR
61

FIG. 6

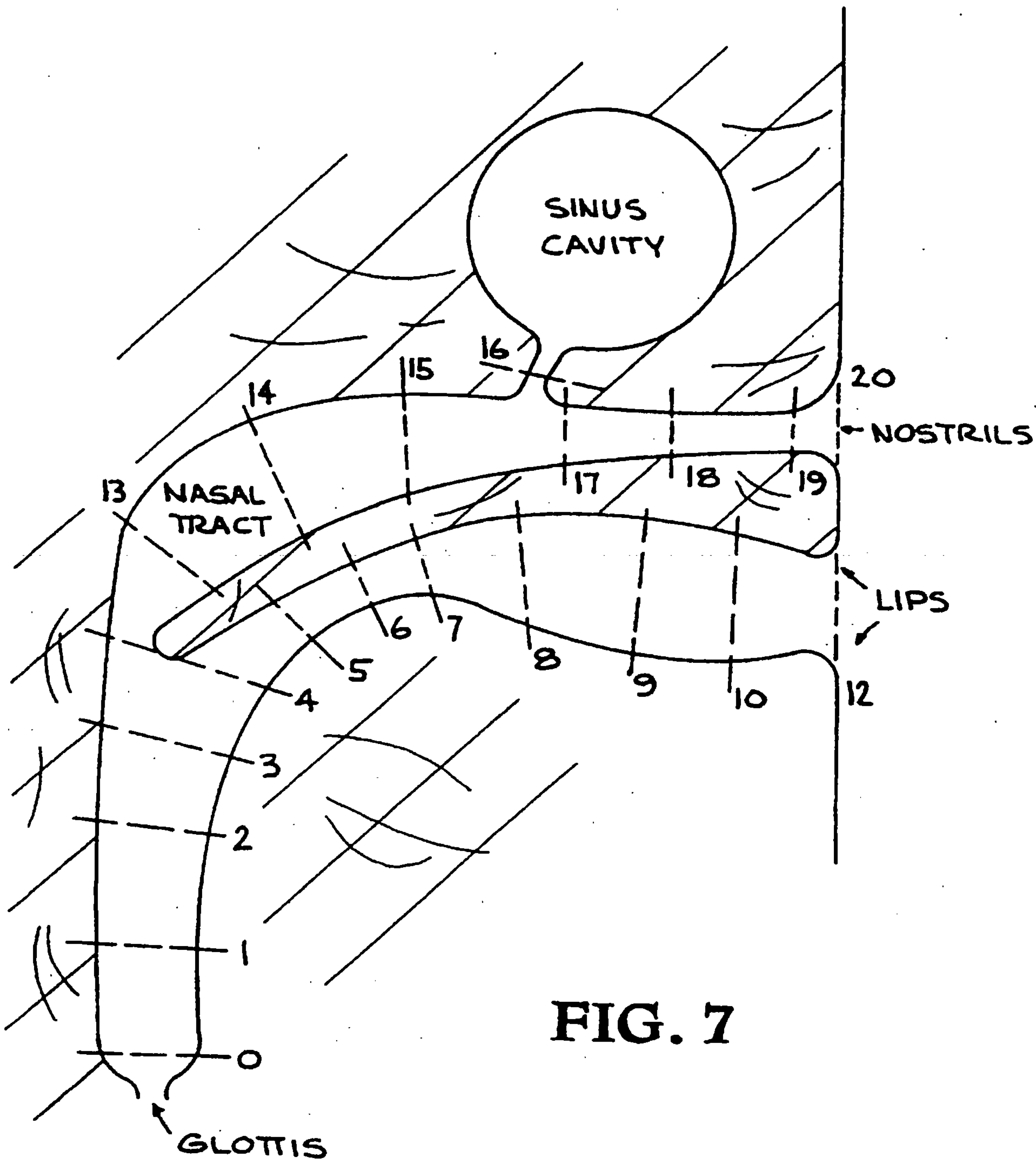

SINUS
CAVITY
16
15
14
20
NOSTRILS
13
NASAL
TRACT
17
18
19
LIPS
6
7
5
4
8
9
10
12
3
2
1
0
GLOTTIS

FIG. 7

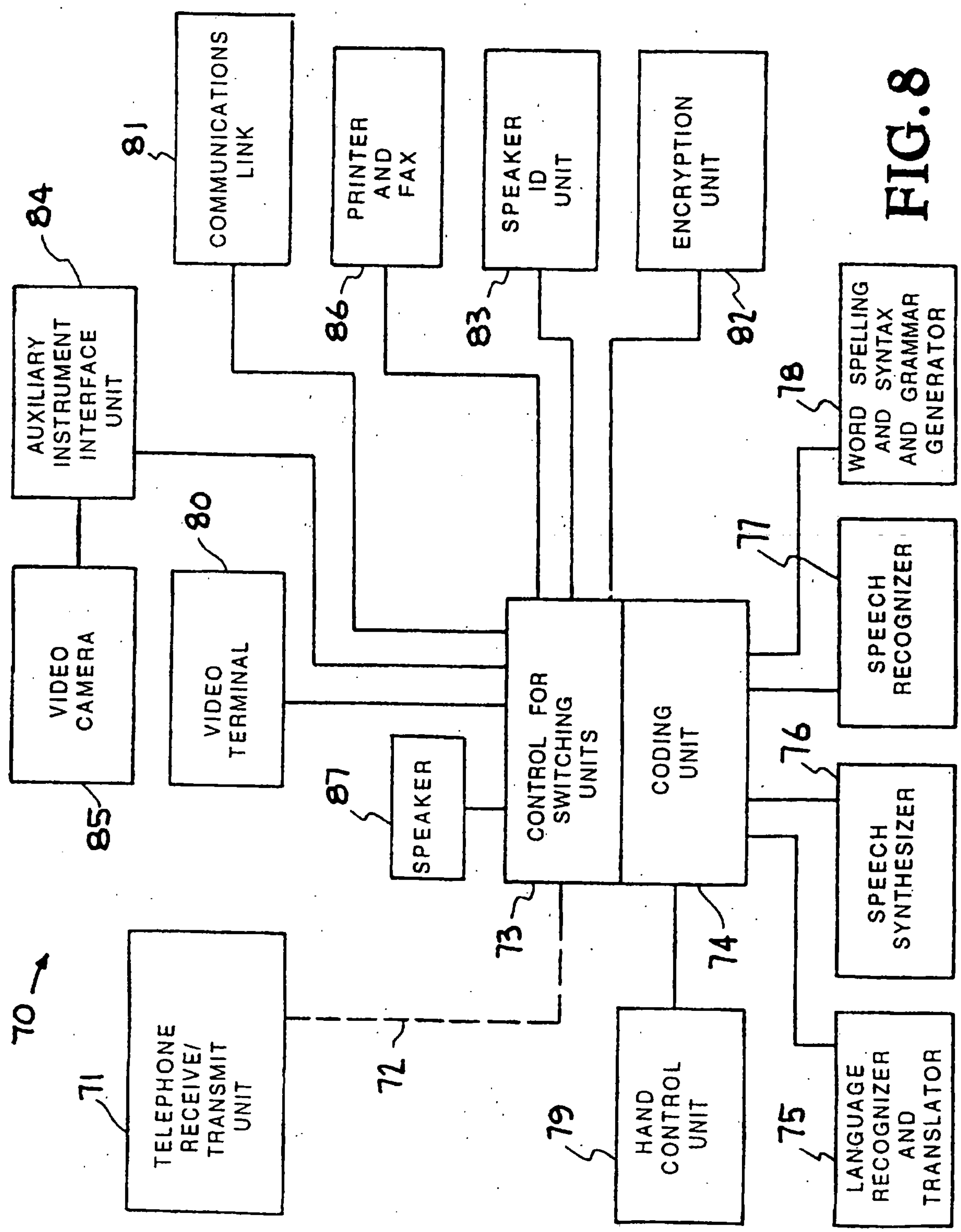

COMMUNICATIONS LINK
81
PRINTER AND FAX
86
SPEAKER ID UNIT
83
ENCRYPTION UNIT
82
AUXILIARY INSTRUMENT INTERFACE UNIT
84
WORD SPELLING AND SYNTAX AND GRAMMAR GENERATOR
78
VIDEO CAMERA
85
VIDEO TERMINAL
80
SPEECH RECOGNIZER
77
CONTROL FOR SWITCHING UNITS
73
CODING UNIT
74
SPEAKER
87
SPEECH SYNTHESIZER
76
TELEPHONE RECEIVE/ TRANSMIT UNIT
71
72
70
HAND CONTROL UNIT
79
LANGUAGE RECOGNIZER AND TRANSLATOR
75

FIG.8

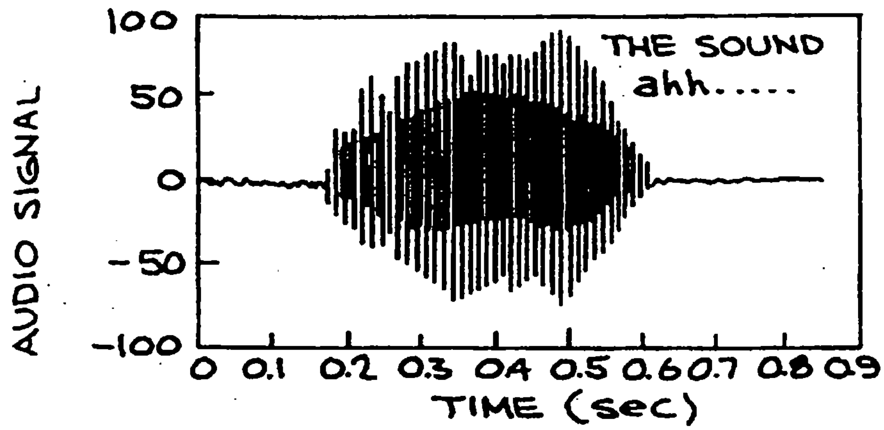


FIG.9A

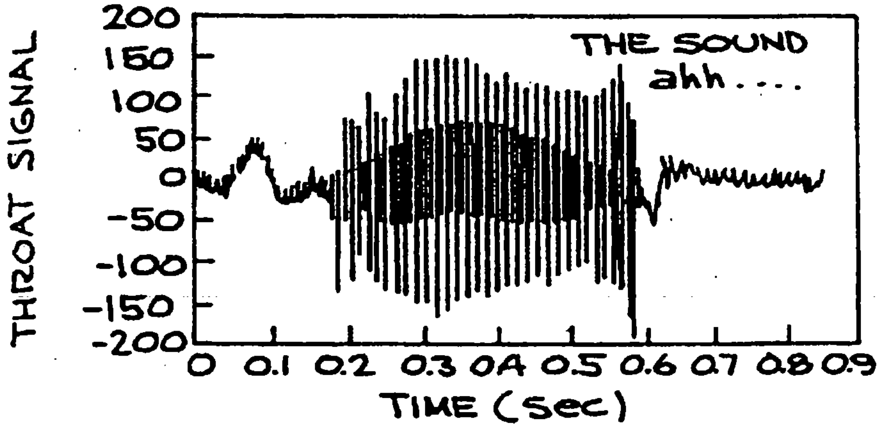


FIG.9B

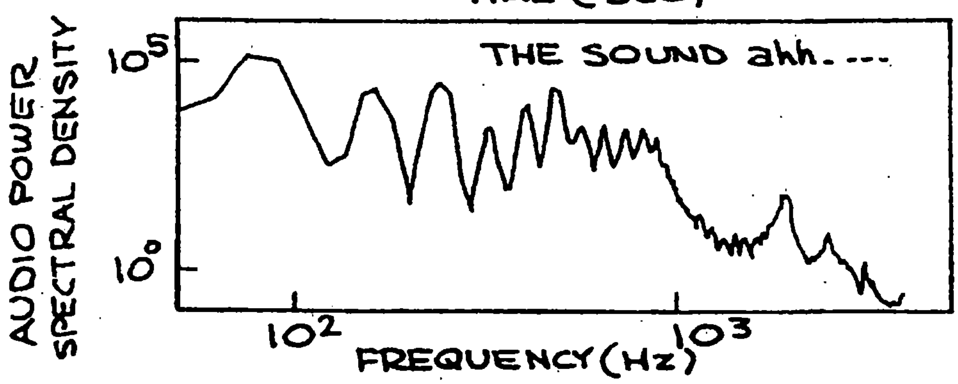


FIG.10A

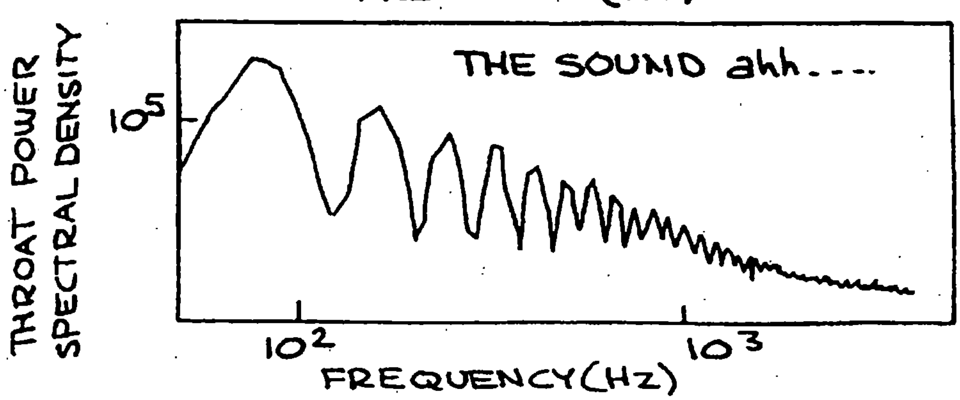


FIG.10B

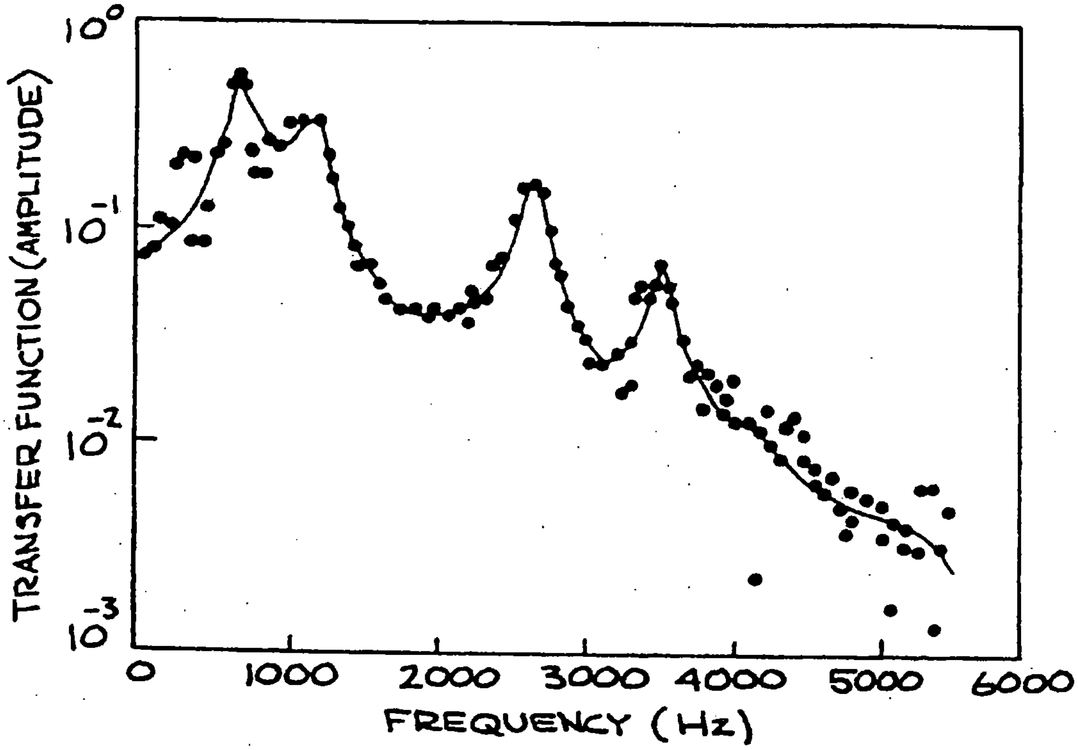


FIG. 11A

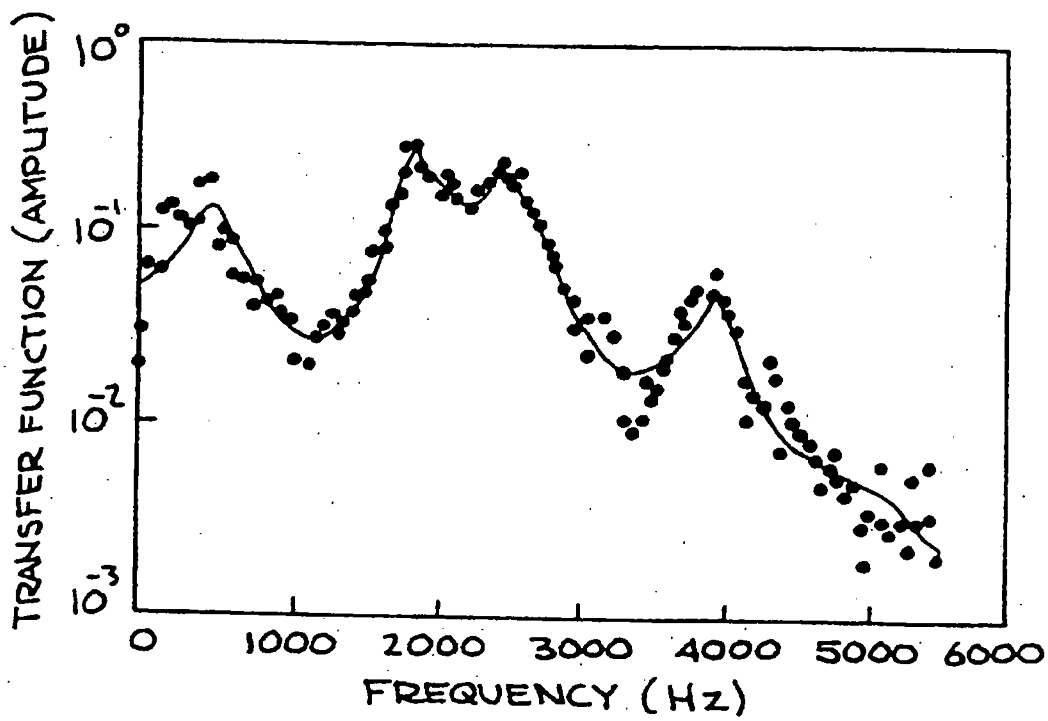


FIG. 11B

SUBSTITUTE SHEET (RULE 26)

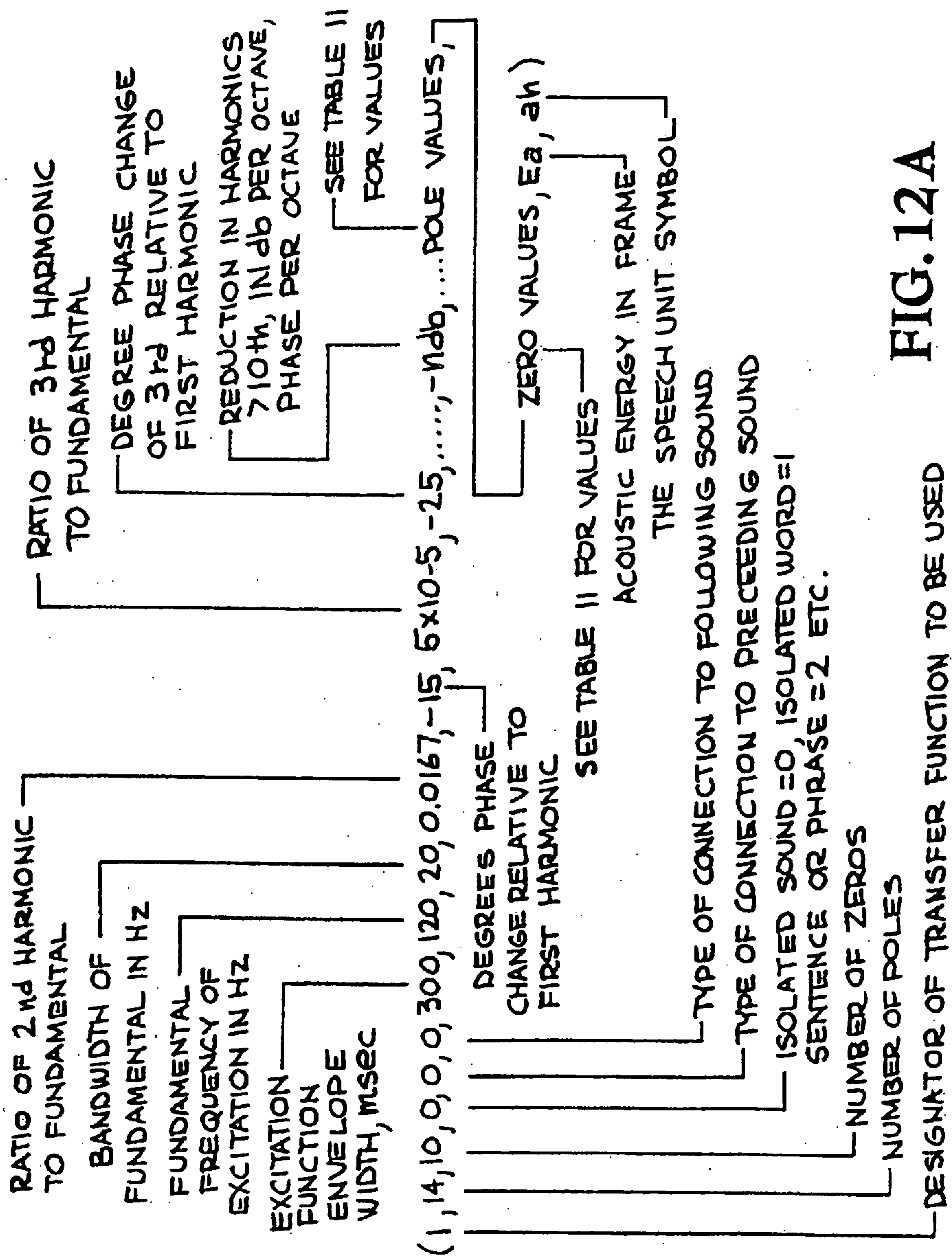

(1, 14, 10, 0, 0, 0, 300, 120, 20, 0.0167, -15, 5x10-5, -25, ..., -ndb, ...POLE VALUES, ZERO VALUES, Ea, ah)
RATIO OF 2nd HARMONIC TO FUNDAMENTAL
BANDWIDTH OF FUNDAMENTAL IN Hz
FUNDAMENTAL FREQUENCY OF EXCITATION IN Hz
EXCITATION FUNCTION ENVELOPE WIDTH, msec
DEGREES PHASE CHANGE RELATIVE TO FIRST HARMONIC
RATIO OF 3rd HARMONIC TO FUNDAMENTAL
DEGREE PHASE CHANGE OF 3rd RELATIVE TO FIRST HARMONIC
REDUCTION IN HARMONICS >10 Hz, IN db PER OCTAVE, PHASE PER OCTAVE
SEE TABLE II FOR VALUES
SEE TABLE II FOR VALUES
ACOUSTIC ENERGY IN FRAME
THE SPEECH UNIT SYMBOL
TYPE OF CONNECTION TO FOLLOWING SOUND
TYPE OF CONNECTION TO PRECEEDING SOUND
ISOLATED SOUND =0, ISOLATED WORD=1 SENTENCE OR PHRASE =2 ETC.
NUMBER OF ZEROS
NUMBER OF POLES
DESIGNATOR OF TRANSFER FUNCTION TO BE USED

FIG. 12A

Z-Plane Zeros	Z-plane Poles
0.4576 + 1.4440i	-0.9289
0.4576 - 1.4440i	-0.6860 + 0.6104i
1.0815 + 0.8980i	-0.6860 - 0.6104i
1.0815 - 0.8980i	-0.4111 + 0.8540i
-0.3252 + 1.2521i	-0.4111 - 0.8540i
-0.3252 - 1.2521i	0.0367 + 0.9656i
-0.8458 + 0.7755i	0.0367 - 0.9656i
-0.8458 - 0.7755i	0.8825
1.1037	0.8754 + 0.3664i
-0.9574	0.8754 - 0.3664i
	0.7405 + 0.6016i
	0.7405 - 0.6016i
	0.2343 + 0.7068i
	0.2343 - 0.7068i

FIG. 12B

coeff. vector a	coeff. vector b
1.0000	0.0055
-2.0451	-0.0350
1.9064	0.0419
-0.9941	-0.0166
0.8417	0.0285
-0.4840	-0.0142
-0.4557	-0.0331
0.8479	-0.0341
-0.1022	-0.0208
-0.3799	
0.1873	

FIG. 12C

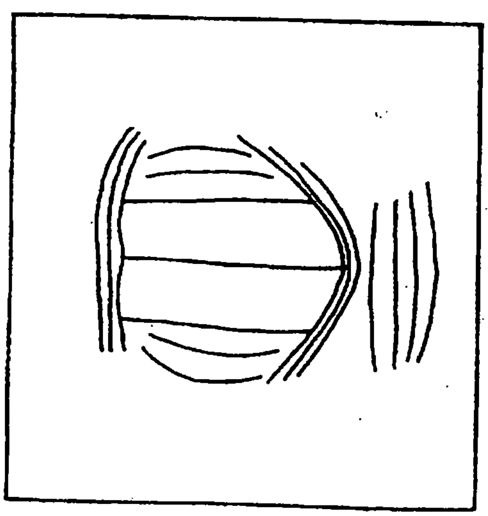


FIG.13A

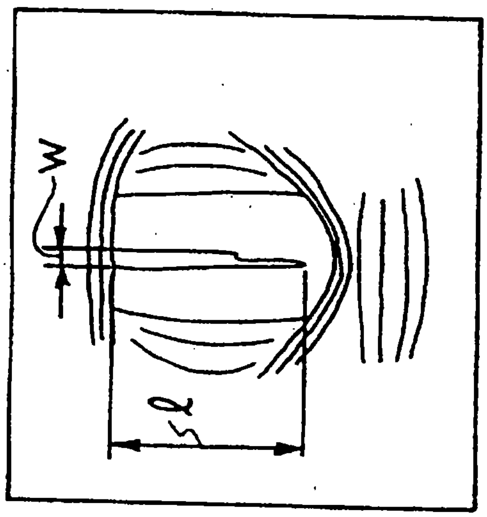


FIG.13B

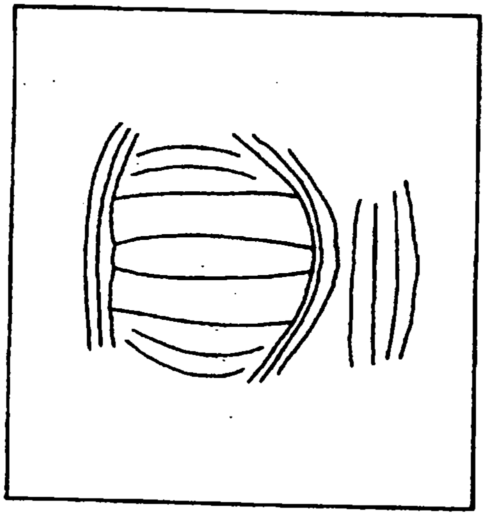


FIG.13C

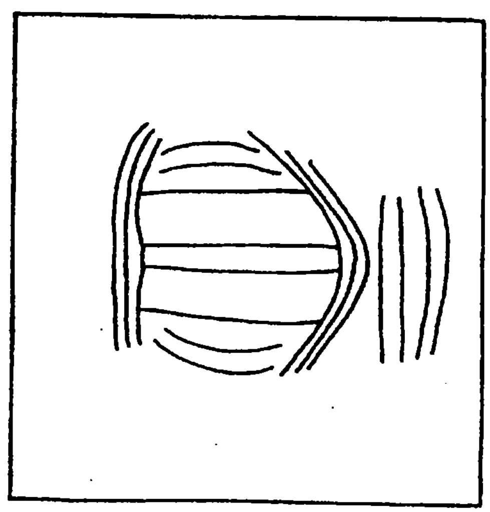


FIG.13D

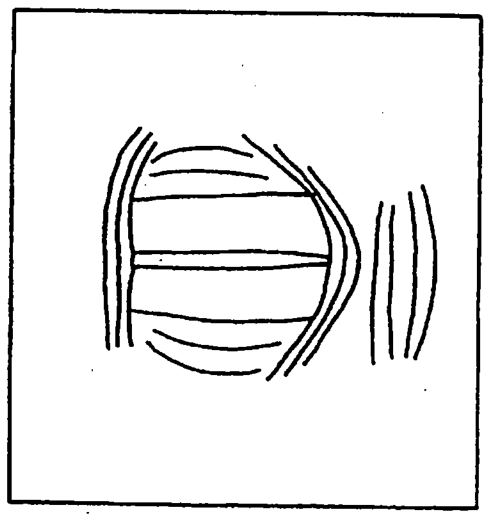


FIG.13E

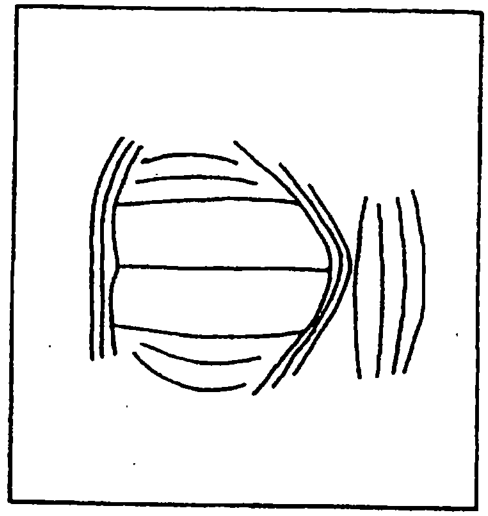


FIG.13F

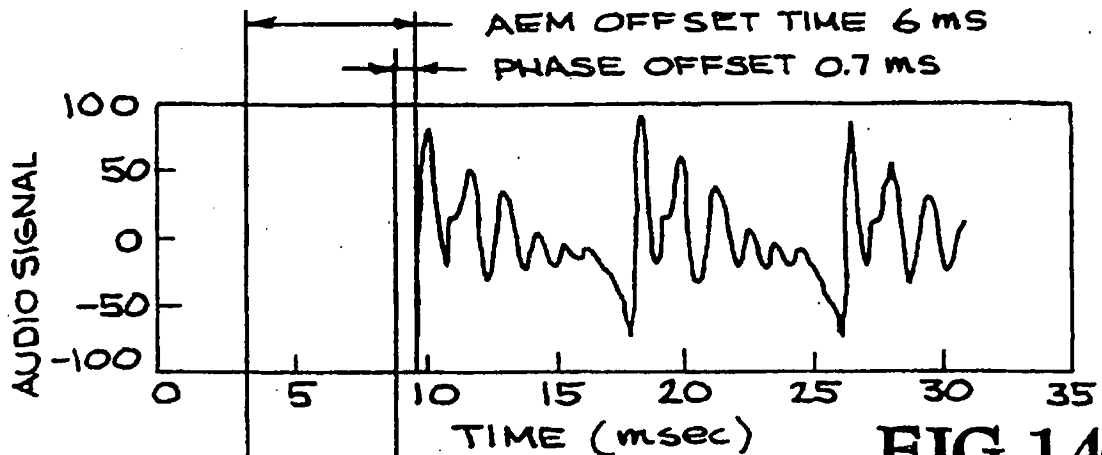


FIG.14A

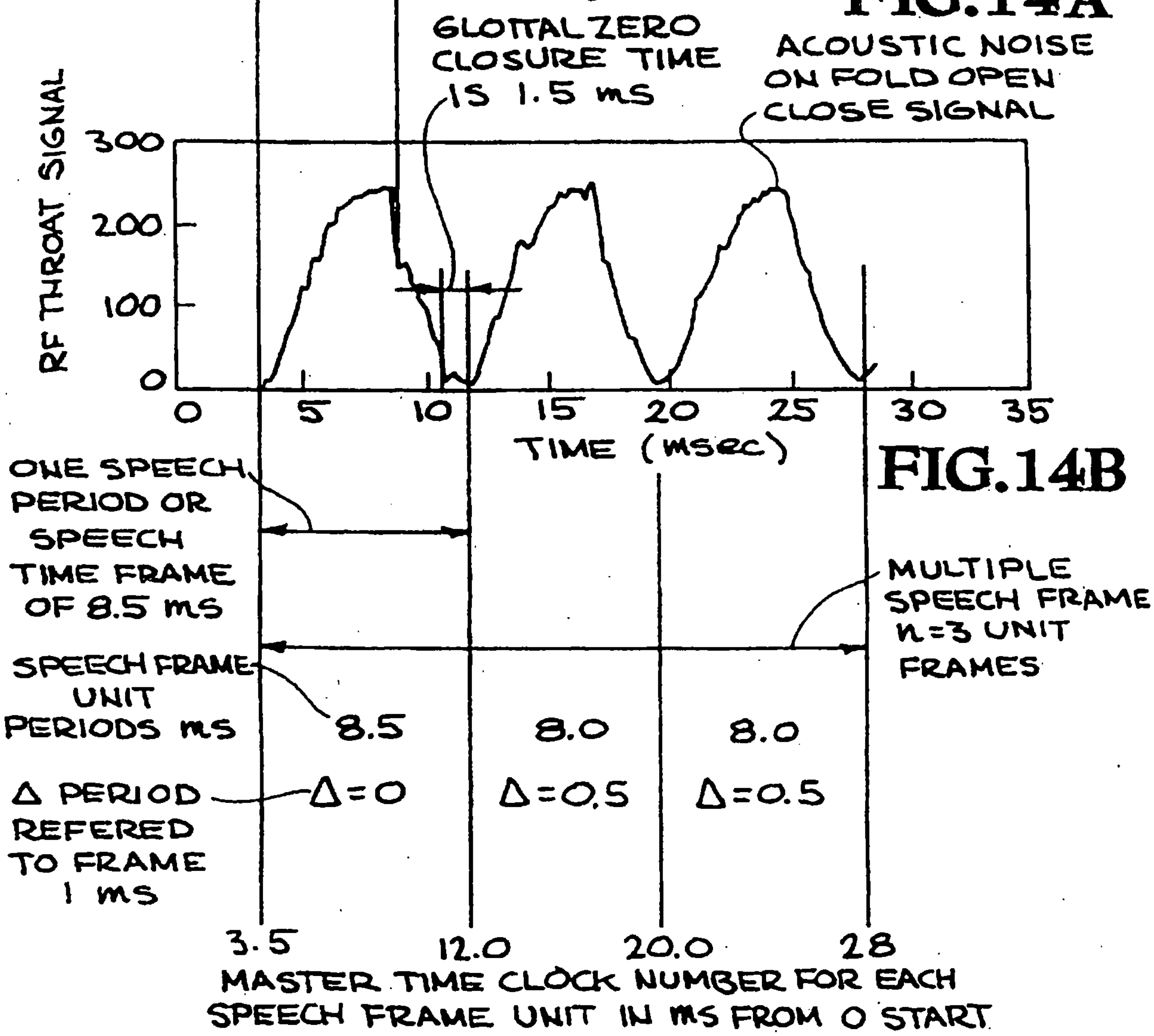


FIG.14B

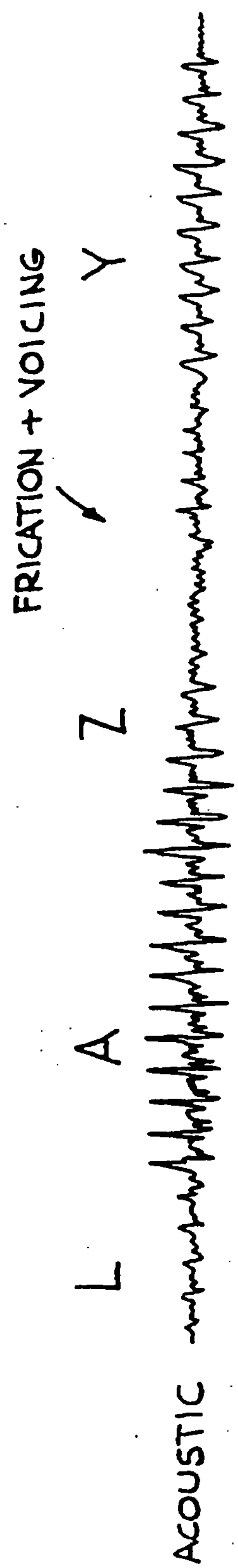


FIG. 15A

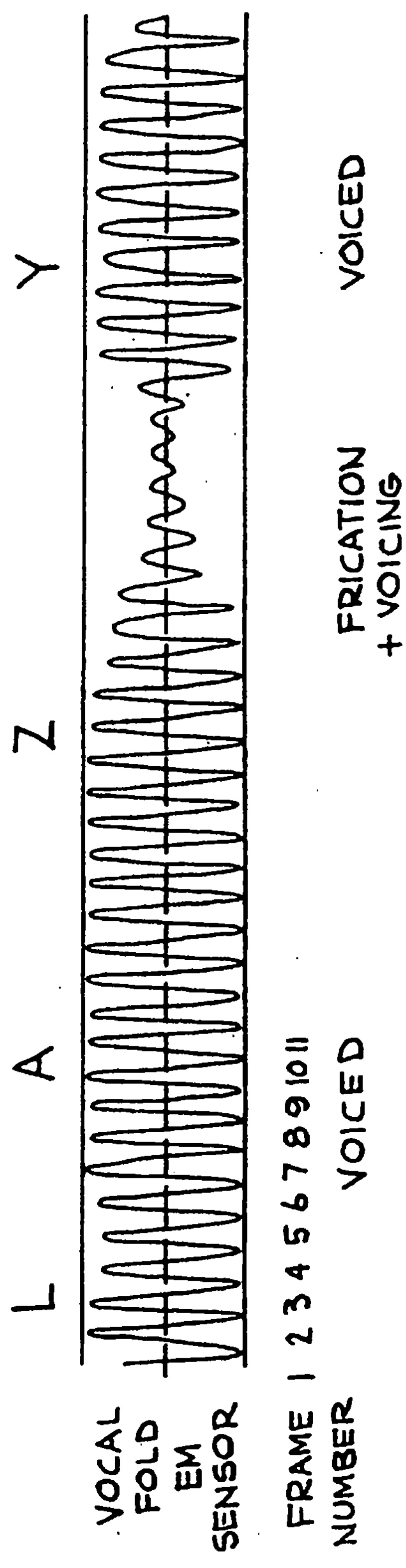


FIG. 15B

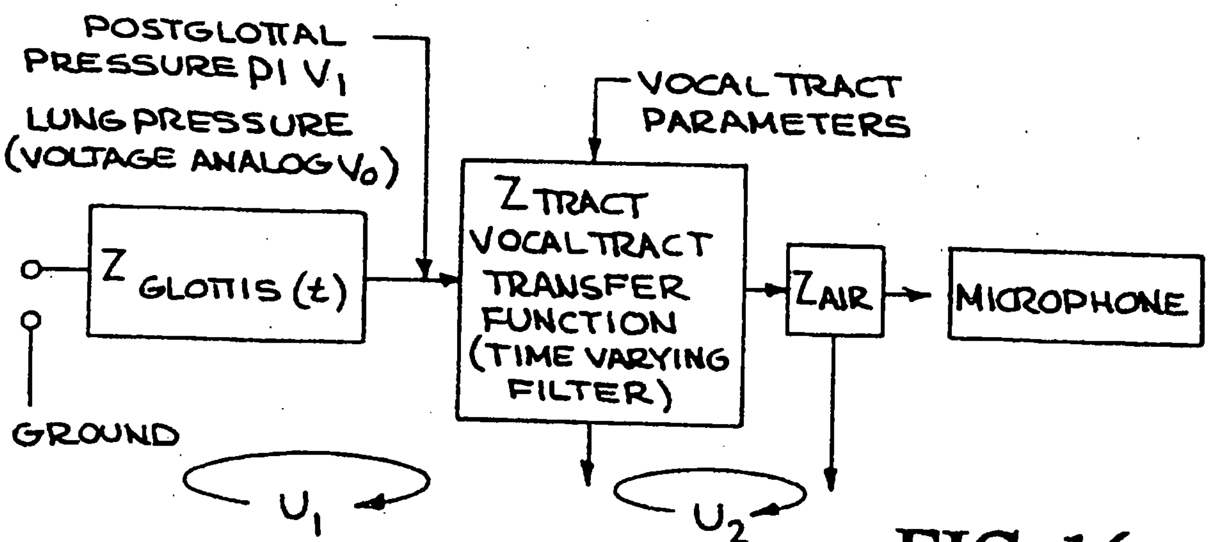


FIG. 16

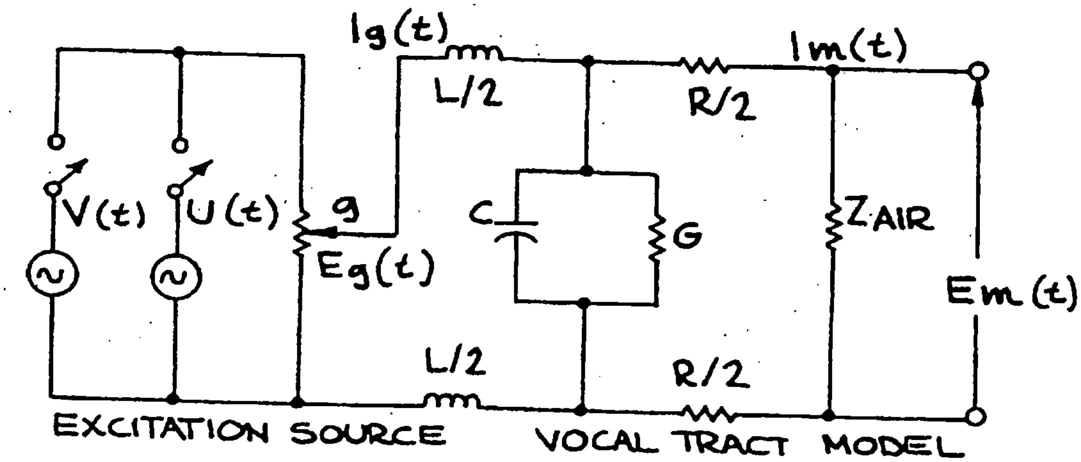


FIG. 17A

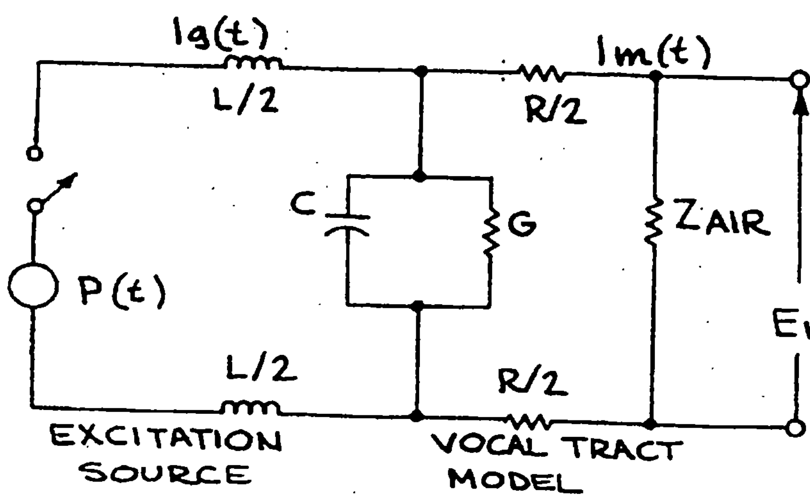


FIG. 17B

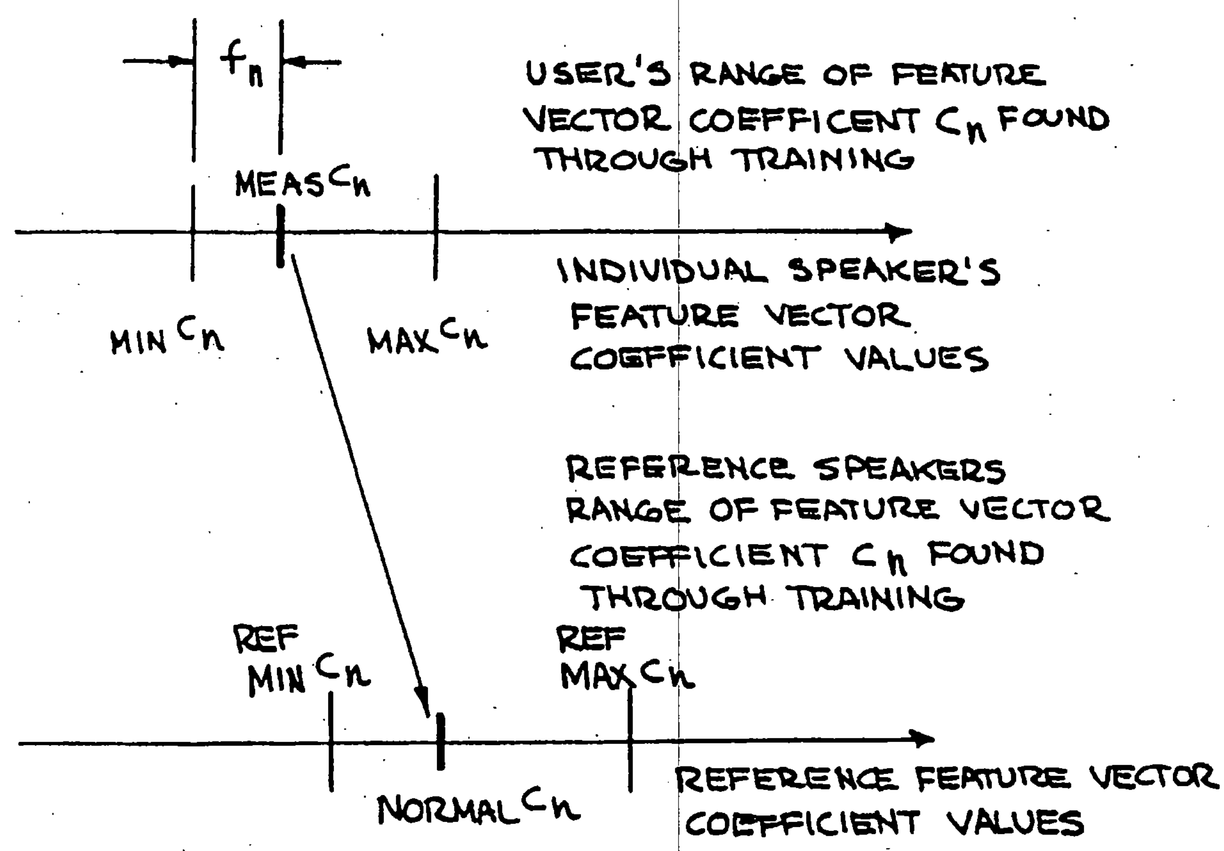

f_n
MEAS C_n
MIN C_n
MAX C_n
USER'S RANGE OF FEATURE VECTOR COEFFICENT C_n FOUND THROUGH TRAINING
INDIVIDUAL SPEAKER'S FEATURE VECTOR COEFFICIENT VALUES
REFERENCE SPEAKERS RANGE OF FEATURE VECTOR COEFFICIENT C_n FOUND THROUGH TRAINING
REF MIN C_n
REF MAX C_n
NORMAL C_n
REFERENCE FEATURE VECTOR COEFFICIENT VALUES

FIG. 18A

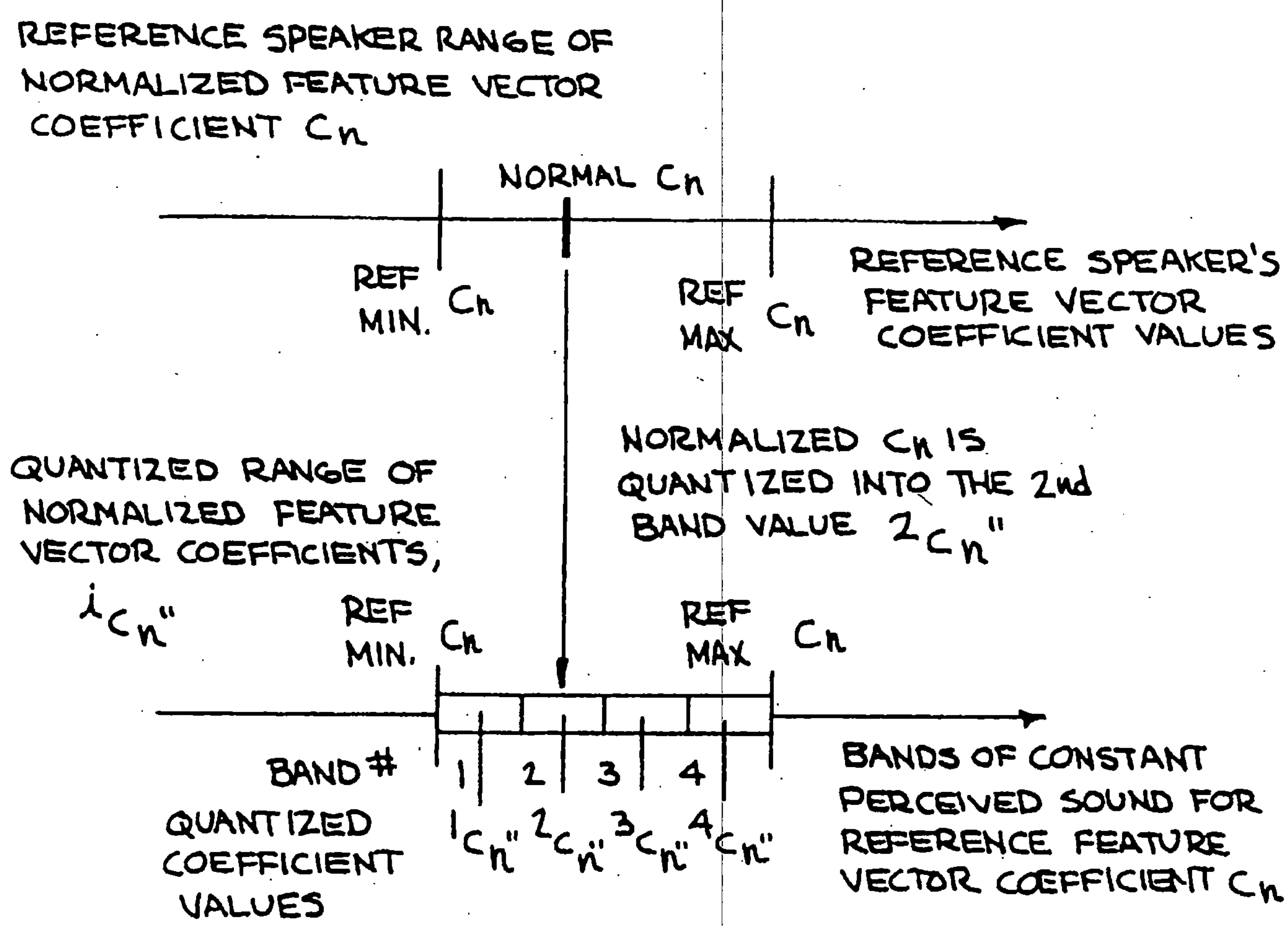

REFERENCE SPEAKER RANGE OF NORMALIZED FEATURE VECTOR COEFFICIENT Cn
NORMAL Cn
REF MIN. Cn
REF MAX Cn
REFERENCE SPEAKER'S FEATURE VECTOR COEFFICIENT VALUES
QUANTIZED RANGE OF NORMALIZED FEATURE VECTOR COEFFICIENTS, iCn"
NORMALIZED Cn IS QUANTIZED INTO THE 2nd BAND VALUE 2Cn"
REF MIN. Cn
REF MAX Cn
BAND #
1
2
3
4
QUANTIZED COEFFICIENT VALUES
1Cn"
2Cn"
3Cn"
4Cn"
BANDS OF CONSTANT PERCEIVED SOUND FOR REFERENCE FEATURE VECTOR COEFFICIENT Cn

FIG. 18B

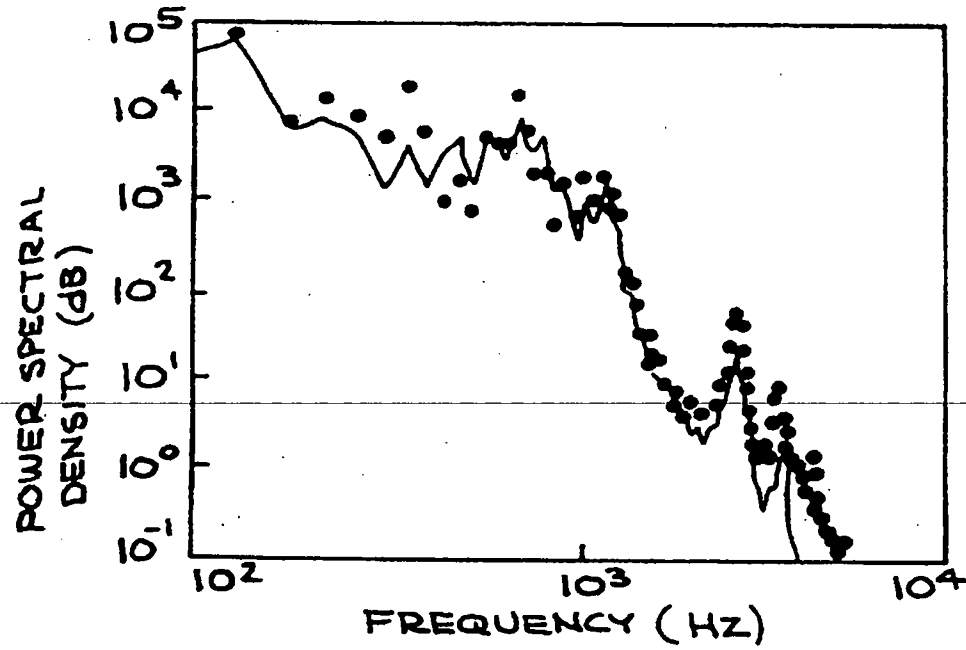

10^5
10^4
10^3
10^2
10^1
10^0
10^{-1}
POWER SPECTRAL DENSITY (dB)
10^2
10^3
10^4
FREQUENCY (Hz)

FIG. 19

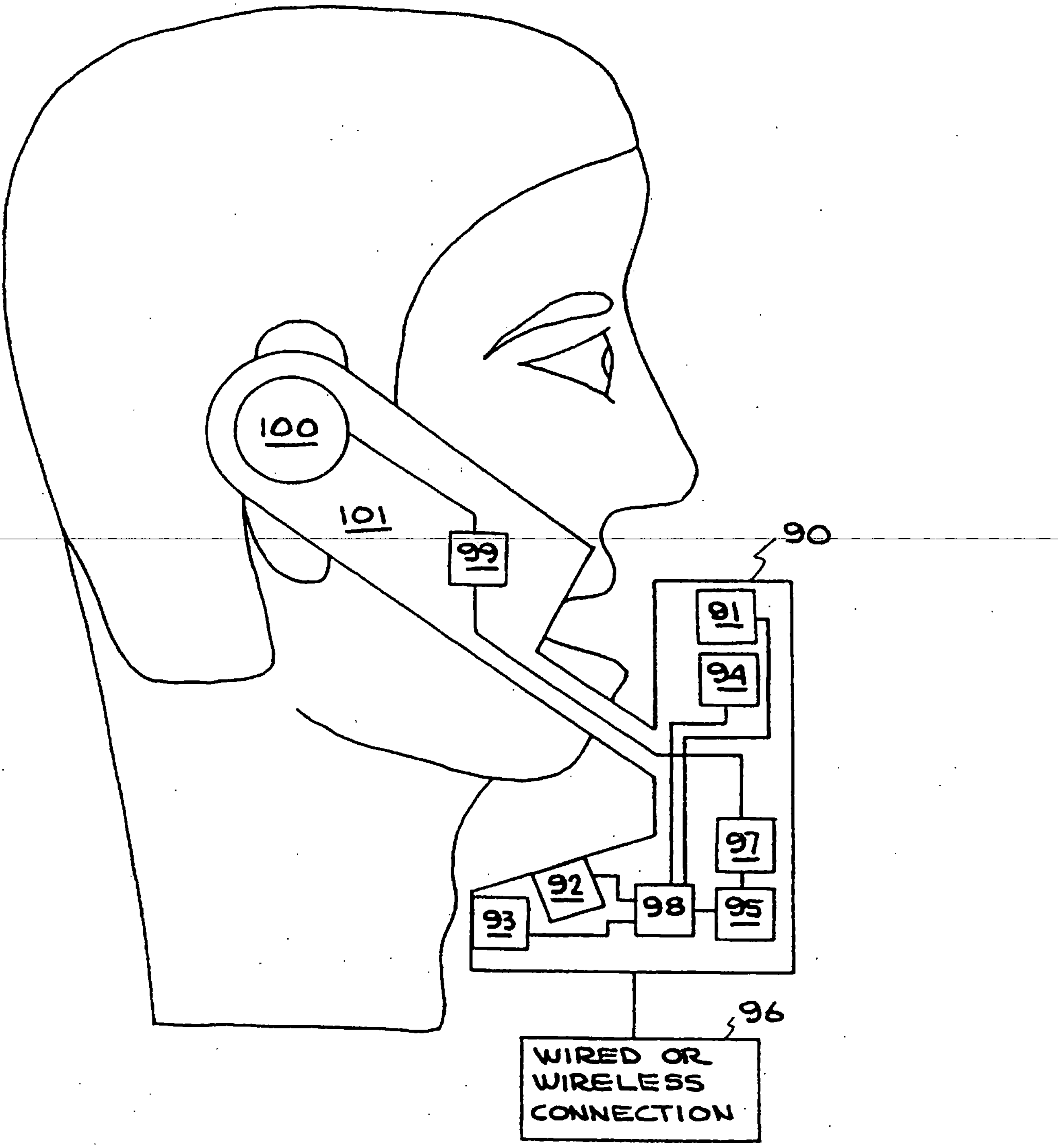

100
101
99
90
81
94
97
92
98
95
93
96
WIRED OR
WIRELESS
CONNECTION

FIG. 20

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US97/01490

A. CLASSIFICATION OF SUBJECT MATTER	
IPC(6) :G10L 9/00 US CL :395/2.1, 2.17 According to International Patent Classification (IPC) or to both national classification and IPC	
B. FIELDS SEARCHED	
Minimum documentation searched (classification system followed by classification symbols) U.S. : 395/2.1, 2.17	
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched	
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) APS, IEEE Proquest	

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A, P	US 5,549,658 (Shannon et al) 27 August 1996, col. 9, line 35 - col. 10, line 66.	1-43

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be part of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 26 MARCH 1997	Date of mailing of the international search report 16.04.1997
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer TARIQ R. HAFIZ Telephone No. (703) 305-9643

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 872 808 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
21.10.1998 Bulletin 1998/43

(51) Int. Cl.[6]: G06K 11/08

(21) Application number: 98106380.3

(22) Date of filing: 07.04.1998

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: 18.04.1997 US 844444

(71) Applicant:
DIGITAL EQUIPMENT CORPORATION
Maynard Massachusetts 01754-1418 (US)

(72) Inventors:
- Waters, Keith
  West Newton, Massachusetts 02165 (US)
- Rehg, James M.
  Arlington, Massachusetts 02174 (US)
- Loughlin, Maria
  Belmont, Massachusetts 02178 (US)
- Kang, Sing Bing
  Cambridge, Massachusetts 02138 (US)

(74) Representative: Betten & Resch
Reichenbachstrasse 19
80469 München (DE)

(54) Method and apparatus for visual sensing of humans for active public interfaces

(57) An active public user interface in a computerized kiosk senses humans visually using movement and color to detect changes in the environment indicating the presence of people. Interaction spaces are defined and the system records an initial model of its environment which is updated over time to reflect the addition or subtraction of inanimate objects and to compensate for lighting changes. The system develops models of the moving objects and is thereby able to track people as they move about the interaction spaces. A stereo camera system further enhances the system's ability to sense location and movement. The kiosk presents audio and visual feedback in response to what it "sees."

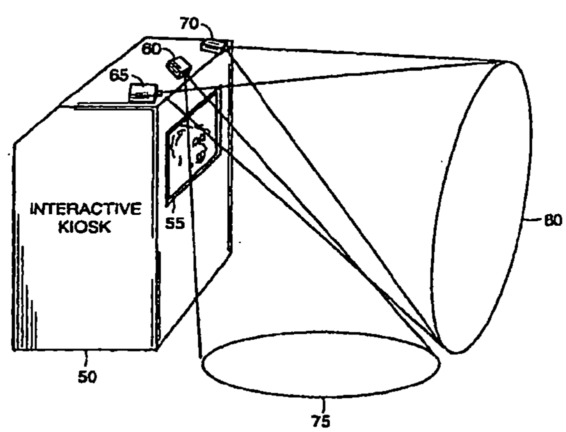


FIG. 2

EP 0 872 808 A1

Description

FIELD OF THE INVENTION

This invention relates generally to computer systems, and more particularly to computerized human-computer interfaces.

BACKGROUND OF THE INVENTION

Computer vision-based sensing of users enables a new class of public multi-user computer interfaces. An interface such as an automated information dispensing kiosk represents a computing paradigm that differs from the conventional desktop environment and correspondingly requires a user interface that is unlike the traditional Window, Icon, Mouse and Pointer (WIMP) interface. Consequently, as user interfaces evolve and migrate off the desktop, vision-based human sensing will play an increasingly important role in human-computer interaction.

Human sensing techniques that use computer vision can play a significant role in public user interfaces for kiosk-like computerized appliances. Computer vision using unobtrusive video cameras can provide a wealth of information about users, ranging from their three dimensional location to their facial expressions, and body posture and movements. Although vision-based human sensing has received increasing attention, relatively little work has been done on integrating this technology into functioning user interfaces.

The dynamic, unconstrained nature of a public space, such as a shopping mall, poses a challenging user interface problem for a computerized kiosk. This user interface problem can be referred to as the public user interface problem, to differentiate it from interactions that take place in a structured, single-user desktop environments. A fully automated public kiosk interface must be capable of actively initiating and terminating interactions with users. The kiosk must also be capable of dividing its resources among multiple users in an equitable manner.

The prior art technique for sensing users as applied in the Alive system is described in "Pfinder: Real-time Tracking of the Human Body," Christopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland, IEEE 1996. Another prior art system is described in "Real-time Self-calibrating Stereo Person Tracking Using 3-D Shape Estimation from Blob Features," Ali Azarbayejani and Alex Pentland, ICPR January 1996.

The Alive system senses only a single user, and addresses only a constrained virtual world environment. Because the user is immersed in a virtual world, the context for the interaction is straight-forward, and, simple vision and graphics techniques can be employed. Sensing multiple users in an unconstrained real-world environment, and providing behavior-driven output in the context of that environment presents more complex vision and graphics problems stemming from the requirement of real world interaction that are not addressed in prior art systems.

The Alive system fits a specific geometric shape model, such as a Gaussian ellipse, to a description representing the human user. The human shape model is referred to as a "blob." This method of describing shapes is generally inflexible. The Alive system uses a Gaussian color model which limits the description of the users to one dominant color. Such a limited color model limits the ability of the system to distinguish among multiple users.

The prior art system, supra, by Azarbayejani uses a self-calibrating blob stereo approach based on a Gaussian color blob model. This system has all of the disadvantages of inflexibility of the Gaussian model. The self-calibrating aspect of this system may be applicable to a desktop setting, where a single user can tolerate the delay associated with self-calibration. In a kiosk setting, it would be preferable to calibrate the system in advance so it will function immediately for each new user.

The prior art systems use the placement of the user's feet on the ground plane to determine the position of the user within the interaction space. This is a reasonable approach in a constrained virtual-reality environment, but this simplistic method is not acceptable in a real-world kiosk setting where the user's feet may not be visible due to occlusion by nearer objects in the environment. Furthermore, the requirement to detect the ground plane may not be convenient in practice because it tends to put strong constraints on the environment.

It remains desirable to have an interface paradigm for a computerized kiosk in which computer vision techniques are used not only to sense users but also to interact with them.

SUMMARY OF THE INVENTION

The problems of the public user interface for computers are solved by the present invention of a computer vision technique for the visual sensing of humans, the modeling of response behaviors, and audiovisual feedback to the user in the context of a computerized kiosk.

The invention, in its broad form, resides in a computerized method and apparatus for interacting with a moving object in a scene observable with a camera, as recited in claims 1 and 10 respectively.

In a preferred embodiment described hereinafter, the kiosk has three basic functional components: a visual sensing component, a behavior module and a graphical/audio module. It has an optional component that contains three dimensional information of the environment, or observed scene. These components interact with each other to produce the effect of a semi-intelligent reaction to user behavior. The present invention is implemented using real-time visual sensing (motion detection, color tracking, and stereo ranging),

and a behavior-based module to generate output depending on the visual input data.

BRIEF DESCRIPTION OF THE DRAWINGS

A more detailed understanding of the invention may be had from the following description of a preferred embodiment, given by way of example, and to be understood with reference to the accompanying drawing wherein:

- FIG. 1 is a block diagram of a public computerized user interface;
- FIG. 2 shows a kiosk and interaction spaces;
- FIG. 3 is shows a block diagram of the kiosk;
- FIG. 4 shows a four zone interaction space;
- FIG. 5 shows a flow diagram of an activity detection program;
- FIG. 6 is a block diagram of a behavior module process; and
- FIG. 7 shows an arrangement for stereo detection of users.

DETAILED DESCRIPTION

Referring now to the figures, FIG. 1 shows a public computer user interface 10. The user interface 10 has a sensing module 15 which takes in information from a real world environment 20, including the presence and actions of users. The information is processed in a behavior module 25 that uses a three dimensional model 30 to determine proper output through a feedback module 35. The three dimensional model 30 of a real world environment 20, also referred to as a scene, includes both metric information and texture that reflect the appearance of the world.

FIG. 2 shows a kiosk 50 with a display screen 55 for the users of the kiosk, and a plurality of cameras 60, 65, 70 which allow the kiosk 50 to detect the presence of the users. Three cameras are shown, but a single camera, or any multiple of cameras may be used. A first camera 60 is aimed at an area on the floor. The "viewing cone" of the first camera 60 is defined to be a first interaction space 75. Second and third cameras 65, 70 are aimed to cover a distance out into the kiosk environment. In the present embodiment of the invention the second and third cameras 65, 70 are aimed out to 50 feet from the kiosk. The space covered by the second and third cameras 65, 70 is a second interaction space 80.

The kiosk 50 includes a visual sensing module 15 which uses a number of computer vision techniques, activity detection, color recognition, and stereo processing, to detect the presence or absence, and the posture of users in the interaction spaces 75, 80. Posture includes attributes such as movement and three dimensional spatial location of a user in the interaction spaces 75, 80. The kiosk digitizes color frames from the cameras that are used by the visual sensing module 15 in the kiosk.

FIG. 3 is a block diagram of the kiosk 50. The kiosk 50 has input devices which include a plurality of cameras 100 coupled to digitizers 105 and output devices which may, for example, include a speaker 110 for audio output and a display screen 115 for visual output. The kiosk 50 includes a memory/processor 120, a visual sensing module 15, a behavior module 25, and a feedback module 35. The kiosk may also include a three dimensional model 30 representative of the scene 20. The visual sensing module 15 includes a detection module 125, a tracking module 130, and a stereo module 135 components which will be more fully described below.

The activity detection module 125 which uses computer vision techniques to detect the presence and movement of users in the interaction spaces of Figure 2. The kiosk 50 accepts video input of the interaction spaces from one or more cameras. In the first embodiment of the invention, the activity detection module 125 accepts video input from a single camera 60 which is mounted so that it points at the floor, as shown in FIG. 2. In operation, the activity detection module 125 examines each frame of the video signal in real-time to determine whether there is a user in the first interaction space 75, and if so, the speed and direction with which the person is moving. The activity detection module sends a message, or notification, to the behavior module every time a moving object enters and exits the first interaction space 75.

The first interaction space 75 is partitioned into one or four zones in which "blobs" are independently tracked. Where a regular camera lens is used, one zone is appropriate. Where a wide-angle or fisheye lens is used, four zones, as shown in FIG. 4, are used. The four zones are defined as a center zone 250, a left zone 255, a right zone 260, and a back zone 265. In the four zone mode, computations for activity detection are performed independently in each zone. The extra computations make the activity detection program more complex but allow more accurate estimation of the velocity at which the user is moving.

When there are four zones in the first interaction space 75, the kiosk is primarily concerned with blobs in the center zone 250, i.e. potential kiosk users. When a blob first appears in the center zone 250, the blob in a peripheral zone from which the center blob is most likely to have originated is selected. The velocity of this source blob is assigned to the center blob. The activity detection program applies standard rules to determine which peripheral zone (Right, Left or Back) is the source of the blob in the center zone 250.

The activity detection module compares frames by finding the difference in intensity of each pixel on the reference frame with the corresponding pixel in a new digitized frame. Corresponding pixels are considered to be "different" if their gray levels differ by more that a first

pre-defined threshold.

The activity detection program distinguishes between a person and an inanimate object, such as a piece of litter, in the first interaction space 75 by looking for movement of the object's blob between successive images. If there is sufficient movement of the object's blob between successive frames, the object is assumed to be animate. There is "sufficient motion" when the number of pixels that differ in successive images is greater that a second threshold.

FIG. 5 shows a flow chart of the operation of the activity detection program. At initialization of the activity detection program, block 400, the first interaction space 75 is empty and the kiosk 50 records a frame of the floor in the first interaction space 75. This initial frame becomes the reference frame 455 for the activity detection program. Approximately every 30 milliseconds, a new frame is digitized, block 400. A comparison, block 405, is then made between this new frame and the reference frame 455. If the new frame is sufficiently different from the reference frame 455 according to the first predefined pixel threshold value, the activity detection module presumes there is a user in the first interaction space 75, block 410. If the new frame is not sufficiently different, the activity detection program presumes that no one is in the first interaction space 75, block 410. If the activity detection program determines that there is a user in the first interaction space 75, the activity detection program sends a message to the behavior module 25, block 420. If the activity detection program determines that there is no person in the first interaction space 75, the behavior module is sent a notification, block 415, and a new frame is digitized, block 400.

If at block 410, the difference is greater than the first predefined threshold, a notification is also provided to the behavior module, block 420. The message indicates that something animate is present in the interaction space 75. At the same time, a frame history log 425 is initialized with five new identical frames which can be the initial frame (of block 400), block 430. A new frame, captured between significant intervals (approximately once every 10 seconds in the present embodiment), block 435, is then compared with each frame in the log to determine if there is a difference above a second threshold, block 440. The second threshold results in a more sensitive reading than the first threshold. If there is a difference above the second threshold, block 445, the frame is added to the frame history, block 430, a five frame-rotating buffer. The steps of blocks 430, 440, and 445 then repeat which indicates that an animate object has arrived. If there is a difference below the second threshold, block 445, the frame is blended with the reference frame, block 450, to create the new reference frame 455. The end result of the activity detection program is that the background can be slowly evolved to capture inanimate objects that may stray into the environment, as well as accommodate slowly changing characteristics such as lighting changes.

If there is a moving object in the first interaction space 75, the activity detection program computes the velocity of that object by tracking, in each video frame, the location of a representative point of the object's blob, or form. The blob position in successive frames is smoothed to attenuate the effects of noise using known techniques such as Kalman filtering. The activity detection program maintains a record of the existence of potential users in the kiosk interaction space 75 based on detected blobs.

Velocity Computation

The activity detection program computes the velocity of users moving in the first interaction space 75 by tracking blob positions in successive frames. Velocity is used to indicate the "intent" of the blob in the first interaction space 75. That is, the velocity is used to determine whether the blob represents a potential user of the kiosk.

Velocity is computed as a change in position of a blob over time. For the velocity calculation, the blob position is defined as the coordinates of a representative point on the leading edge of the moving blob. When there is only one zone in the interaction space, the representative point is the center of the front edge of the blob. When there are four zones in the interaction space, the representative point differs in each zone. In the center and back zones, the point is the center of the front edge of the blob 252, 267. In the left zone, the point is the front of the right edge of the blob 262. In the right zone, the point is the front of the left edge of the blob 257. The velocities of blobs are analyzed independently in each zone.

Behavior module

The behavior module 25, shown in FIG. 6, uses the output of the visual module 15 as well as a priori information such as the three dimensional model of the environment 30 to formulate actions. The behavior module 25 uses a set of rules (with the potential for learning from examples) as a means of reacting to user behavior in a manner that can be perceived as being intelligent and engaging. The mechanism for reacting to external visual stimuli is equivalent to transitioning between different states in a finite state machine based on known (or learnt) transition rules and the input state. As a simple example, the behavior module 25 can use the output of the detection module 125 to signal the feedback module 35 to acknowledge the presence of the user. It can take the form of a real time talking head in the display screen 55 saying "Hello." Such a talking head is described in "An Automatic Lip-Synchronization Algorithm for Synthetic Faces," Keith Waters and Tom Levergood, Proceedings of the Multimedia ACM Conference, September 1994, pp. 149 - 156. In a more complicated example, using the output of the stereo module 135

(which yields the current three dimensional location of the user/s), the behavior module 25 can command the talking head to focus attention on a specific user by rotating the head to fixate on the user. In the case of multiple users, the behavior module 25 can command the talking head to divide its attention amongst these users. Heuristics may be applied to make the kiosk pay more attention to one user than the other (for example, based on proximity or level of visual activity). In another example, by using both the stereo module 135 and three dimensional world information 30, the behavior module 25 can generate directional information, either visually or orally, to the user based on the user's current three dimensional location.

Color Blob

Color blobs are used to track the kiosk users as they move about the interaction space. The distribution of color in a user's clothing is modeled as a histogram in the YUV color space. A color histogram detection algorithm used by the present invention is described in the context of object detection in "Color Indexing" by Michael J. Swain and Dana H. Ballard, International Journal of Computer Vision, 7:1, 1991, pp. 11 - 32. In the present invention, the color histogram method is used for user tracking and is extended to stereo localization.

Given a histogram model, a histogram intersection algorithm is used to match the model to an input frame. A back projection stage of the algorithm labels each pixel that is consistent with the histogram model. Groups of labeled pixels form color blobs. A bounding box and a center point are computed for each blob. The bounding box and the center point correspond to the location of the user in the image. The bounding box is an x and y minimum and maximum boundary of the blob. The color blob model has advantages for user tracking in a kiosk environment. The primary benefit is that multiple users can be tracked simultaneously, as long as the users are wearing visually distinct clothing. The histogram model can describe clothing with more than one dominant color, making it a better choice than a single color model. Histogram matching can be done very quickly even for an NTSC resolution image (640 by 480 pixels), whereby a single user may be tracked at 30 frames per second. Color blobs are also insensitive to environmental effects. Color blobs can be detected under a wide range of scales, as the distance between the user and the camera varies. Color blobs are also insensitive to rotation and partial occlusion. By normalizing the intensity in the color space, robustness to lighting variations can be achieved. The center locations, however, of detected color blobs are significantly affected by lighting variation. Use of color for tracking requires a reference image from which the histogram model can be built. In the architecture of the present embodiment of the invention, initial blob detection is provided by the activity detection module, which detects moving objects in the frame. The activity detection module assumes that detected blobs correspond to upright moving people, and samples pixels from the central region of the detected blob to build the color histogram model.

Stereo

Through stereo techniques, true three dimensional information about user location can be computed from cameras in an arbitrary position relative to the scene. Stereo techniques require frames from two or more cameras be acquired concurrently, as shown in FIG. 7. This is a known method for computing detailed descriptions of scene geometry. In a classical approach, frames acquired from two cameras are processed and the correspondences between pixels in the pair of frames are determined. Triangulation is used to compute the distance to points in the scene given correspondences and the relative positions of the cameras. In the classical approach, a high level of detail retires excessive computational resources. The method of the present embodiment is based on a simpler, object-based version of the classical stereo technique. Moving objects are tracked independently using color or motion blobs in images obtained from synchronized cameras. Triangulation on the locations of the moving objects in separate views is used to locate the subjects in the scene. Because tracking occurs before triangulation, both the communication and computational costs of dense stereo fusion are avoided.

The triangulation process is illustrated in Figure 7. Given the position of a blob 700 in a first camera image 702, the position of the user 705 is constrained to lie along a ray 710 which emanates from a first camera 715 through the center of the blob 700 and into the scene. Given the position of a second blob 712 in a second camera image 720, the position of the user 705 is constrained to lie along a second ray 725. The user 705 is located at the intersection of the first ray 710 and the second ray 725 in the scene. In actual operation, noise in the positions of the blobs 700, 712 makes it unlikely that the two rays 710, 725 will intersect exactly. The point in the scene where the two rays 710, 725 are closest is therefore chosen as the three dimensional location of the user 705.

In a preferred embodiment of the kiosk system, a pair of verged cameras with a six foot baseline, i.e. separation between the cameras, is used. The stereo approach depends on having calibrated cameras for which both the internal camera parameters and relationship between camera coordinate systems are known. A standard non-linear least squares algorithm along with a calibration pattern to determine these parameters off-line are used.

Camera synchronization is achieved by ganging the external synchronization inputs of the cameras

together. Barrier synchronization is used to ensure that the blob tracking modules that process the camera images begin operation at the same time. Synchronization errors can have a significant effect on conventional stereo systems, but blobs with large size and extent make stereo systems much more robust to these errors.

It is to be understood that the above-described embodiments are simply illustrative of the principles of the invention. The present invention has been described in the context of a kiosk however alternative embodiments could be automated teller machines (ATMs), advanced multimedia TV, or office desk computers. Various and other modifications and changes may be made by those skilled in the art which will embody the principles of the invention and fall within the scope thereof.

Claims

1. A computerized method for interacting with a moving object or person in a scene observable with a camera, comprising the steps of:

   determining a posture of the moving object by comparing successive frames of the scene;
   outputting information which can be sensed by the moving object depending on the posture of the object as determined from the comparison of the successive frames.

2. The method of claim 1, wherein the posture of the moving object includes a position of the moving object, wherein further the position is determined in three dimensional space, and multiple cameras are used to observe the scene.

3. The method of claim 1, wherein the scene includes a plurality of moving objects, the method including observing dominant colors of the plurality of moving objects to interact independently with any of the moving objects.

4. The method of claim 1, wherein the outputted information includes audible and visible signals, further comprising:

   displaying a talking head on a display terminal, the method including controlling the orientation of the talking head depending on the posture of the moving object.

5. The method of claim 4, wherein the step of synchronizing audible signals with the orientation of the talking head and dependent on the posture of the moving object.

6. The method of claim 1 further comprising:

   repeatedly storing a previous frame of the scene in a buffer if a difference between the previous frame and a next frame is greater than a predetermined value;
   determining the posture of the moving object by analyzing the frames stored in the buffer.

7. A computerized apparatus for interacting with a moving object or person in a scene observable with a camera, comprising:

   means for determining a posture of the moving object by comparing successive frames of the scene;
   means for outputting information which can be sensed by the moving object depending on the posture of the object as determined from the comparison of the successive frames.

8. A computerized interface for interacting with people, comprising:

   a camera measuring a region of an arbitrary physical environment as a sequence of images; and
   means for detecting a person in the region from the sequence of images to identify the person as a target for interaction.

9. The interface of Claim 8, further comprising:

   means for rendering audio and visual information directed at the detected person, further comprising:

      means for determining a velocity of the person in the region; and wherein a content of the rendered audio and video information depends on the velocity of the person, wherein the means for rendering includes a display system displaying an image of a head including eyes and a mouth with lips, the display system directing an orientation of the head and a gaze of the eyes at the detected person while rendering the audio information synchronized to movement of the lips so that the head appears to look at and talk to the person.

10. The interface of Claim 9, further comprising:

   means for determining a position and an orientation of the person in the region relative to a position of the camera, further comprising:

      means for rendering audio and video information directed at the detected person, a content of the rendered information

depending upon the determined position and the determined orientation of the person in the region.

11. The interface of Claim 8, further comprising:

a memory, coupled to the means for detecting, storing data representing a three-dimensional model of the physical environment for determining a position of the person in the region relative to objects represented in the three-dimensional model, further comprising:

means for rendering audio and video information, a content of the rendered information depending upon the determined position of the person.

12. The interface of Claim 8, wherein the sequence of images includes a reference image and a target image, each image being defined by pixels, the pixels of the reference image having a one-to-one correspondence to the pixels of the target image; and further comprising:

means for comparing the reference image to the target image to identify a group of adjacent pixels in the reference image that are different from the corresponding pixels in the target image, the identified group of pixels representing the person, wherein the means for comparing compares an intensity of each pixel of the reference image to an intensity of each corresponding pixel in the target image, and the means for detecting detects the presence of the person in the region when the intensities of at least a pre-defined number of the pixels of the reference image differ from the intensities of the corresponding pixels of the target image.

13. The interface of Claim 12 further comprising:

means for blending the target image with the reference image to generate a new reference image when less than a pre-defined number of the pixels of the reference image differ from the corresponding pixels of the target image.

14. The interface of Claim 8 further comprising:

a second camera spaced apart from the other camera, the second camera measuring the region as a second sequence of images, further comprising:

means for determining an approximate three-dimensional position of the person in the region from the sequences of images of the cameras.

15. The interface of Claim 8 further comprising:

means for rendering audio and visual information, the rendered audio and video information interacting in turn with a plurality of detected persons.

16. The interface of Claim 8, wherein the sequence of images includes a reference image and a target image, each image being defined by pixels, the pixels of the reference image having a one-to-one correspondence to the pixels of the target image; and further comprising:

means for comparing the reference image to the target image to identify a plurality of groups of adjacent pixels in the reference image that are different from the corresponding pixels in the target image, each identified group of pixels representing one of a plurality of detected persons.

17. The interface of Claim 16 further comprising:

means for determining a distribution of colors in each of the group of pixels, each color distribution uniquely identifying one of the plurality of persons, further comprising:

means for concurrently tracking movements of each person independently in the region by the color distribution that uniquely identifies that person.

18. A computerized interface for interacting with people, comprising:

a camera measuring a region of an arbitrary physical environment as a sequence of images; and
means for rendering audio and video information directed at a person detected in the region from the sequence of images to interact with the person.

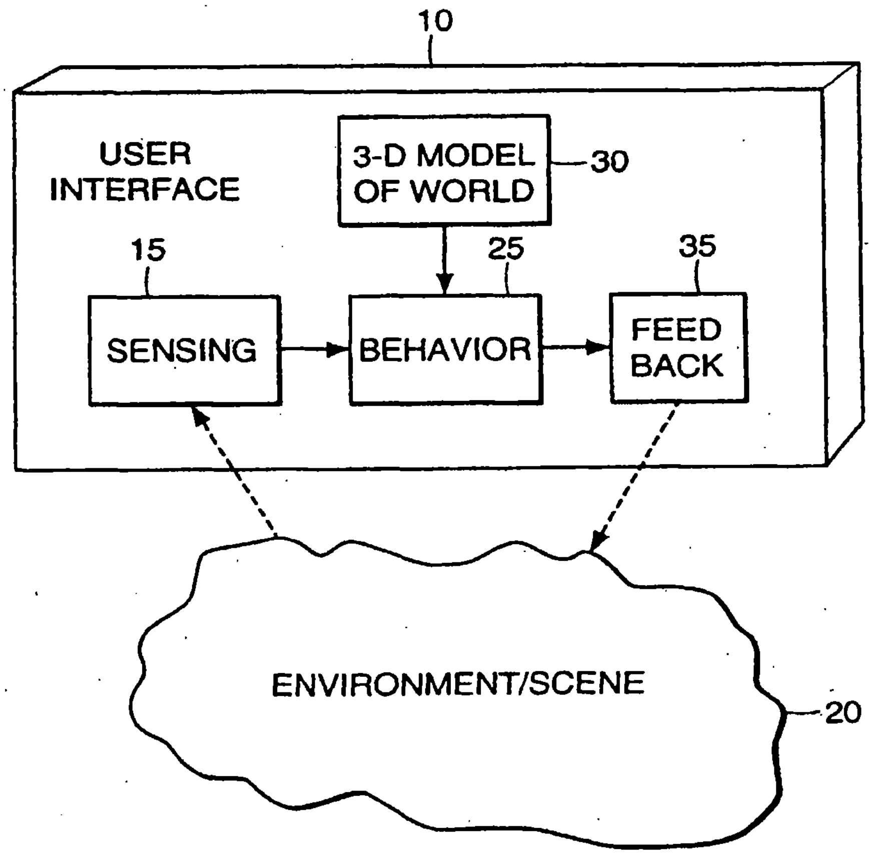

10
USER
INTERFACE
3-D MODEL
OF WORLD
30
15
25
35
SENSING
BEHAVIOR
FEED
BACK
ENVIRONMENT/SCENE
20

FIG. 1

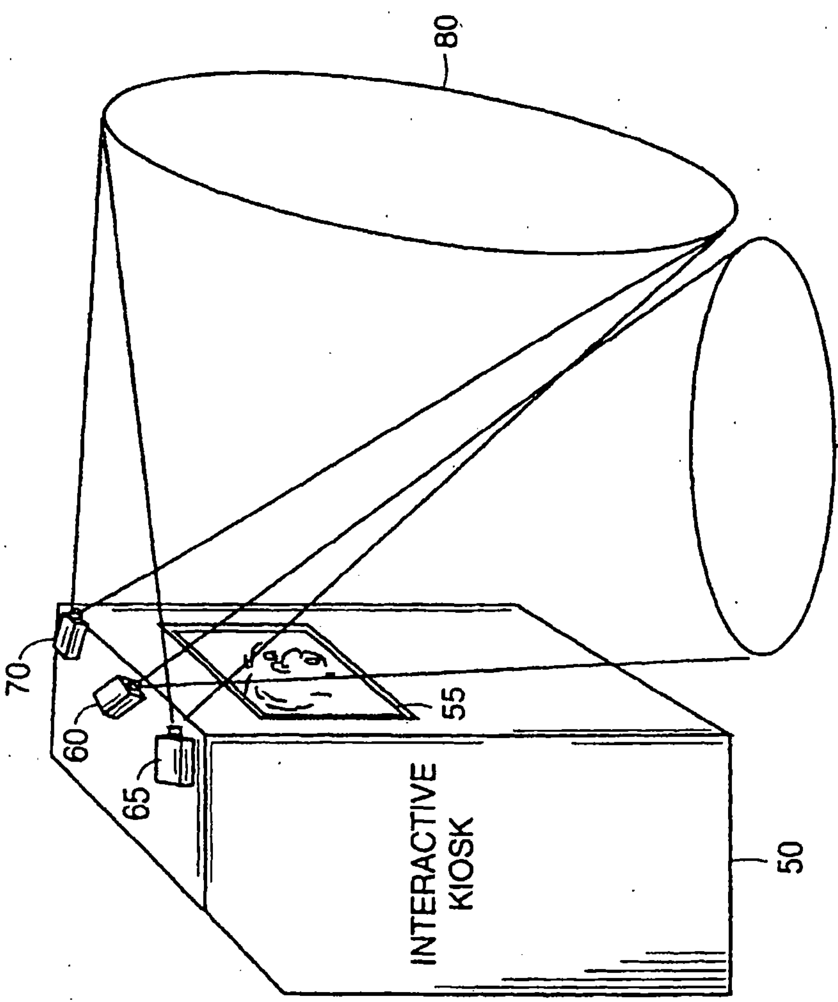


FIG. 2

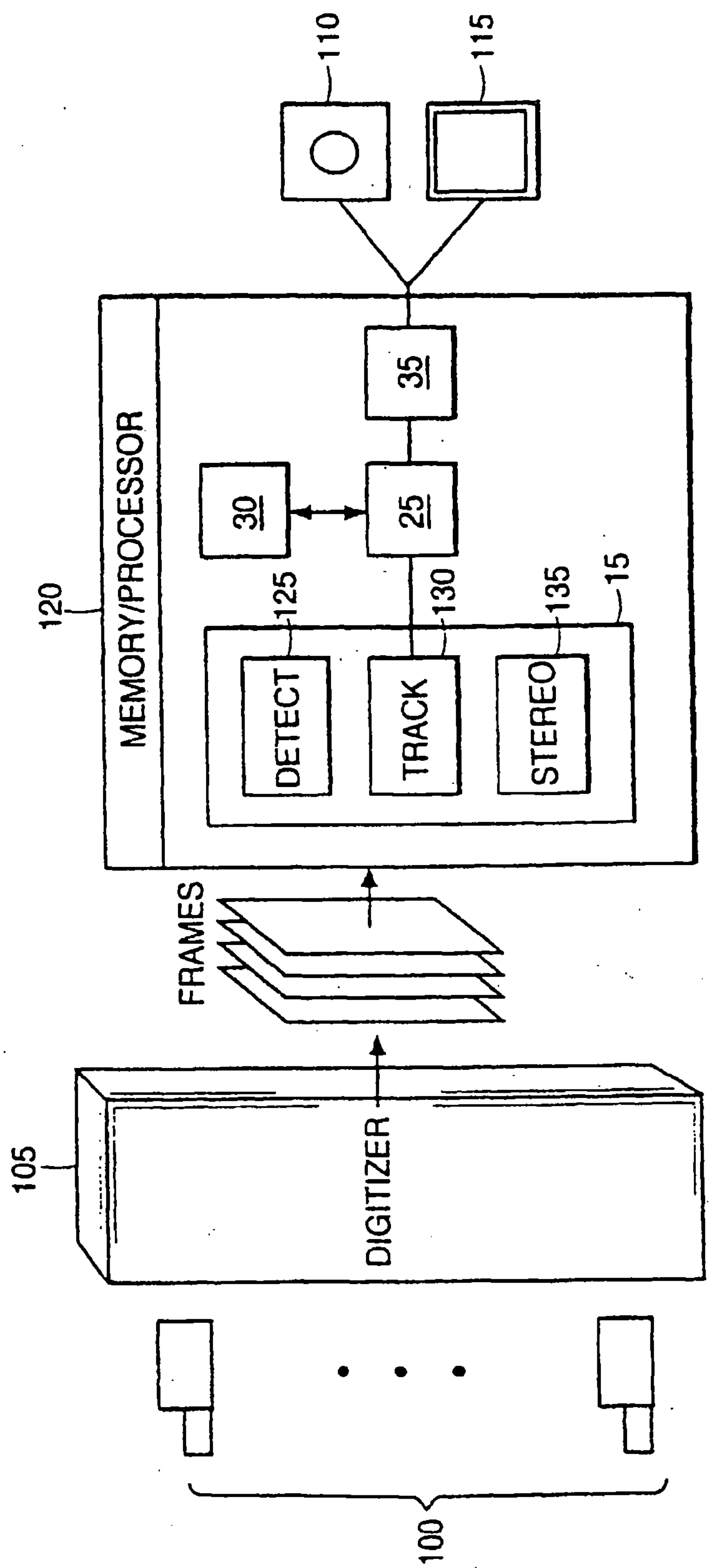

110
115
MEMORY/PROCESSOR
120
35
30
25
125
130
135
15
DETECT
TRACK
STEREO
FRAMES
DIGITIZER
105
100

FIG. 3

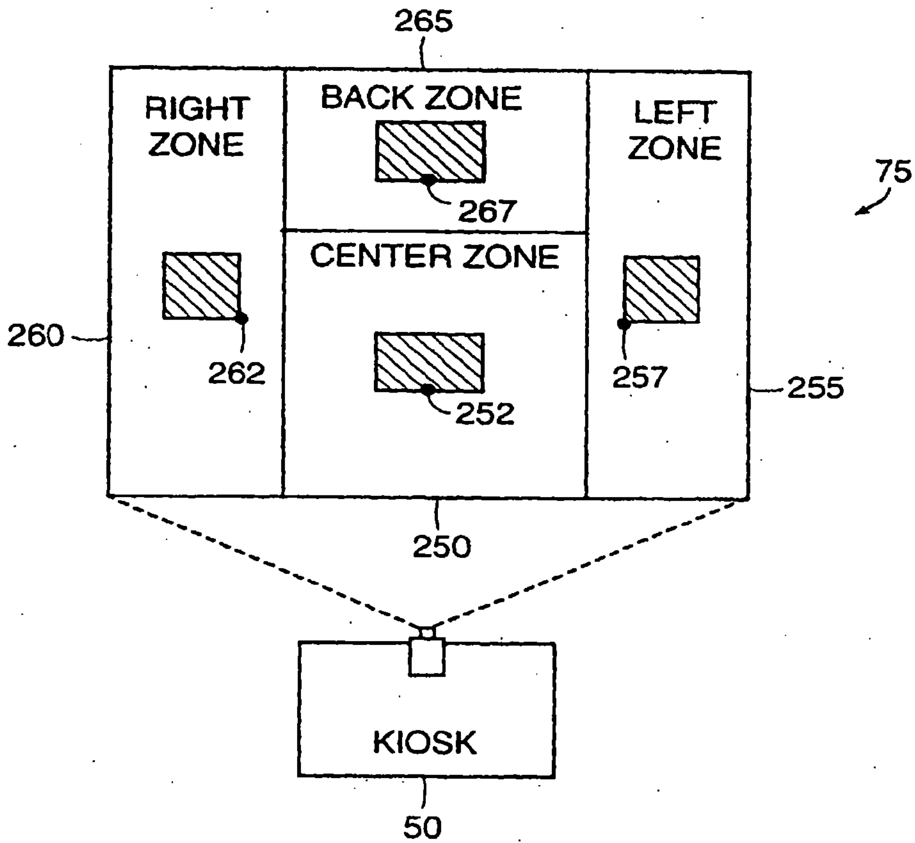

265
RIGHT ZONE
BACK ZONE
LEFT ZONE
75
267
CENTER ZONE
260
262
252
257
255
250
KIOSK
50

FIG. 4

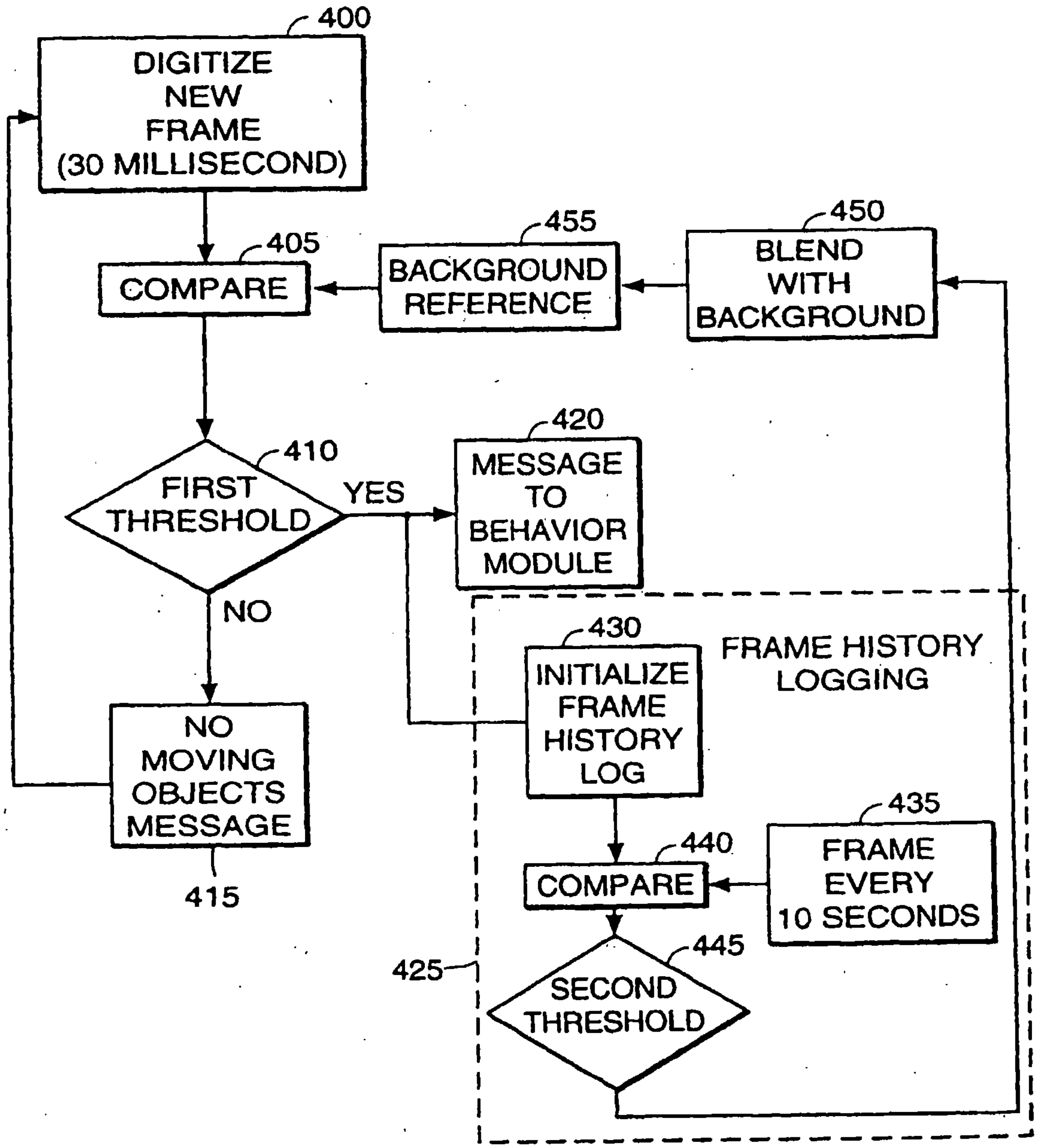

400
DIGITIZE NEW FRAME (30 MILLISECOND)
405
COMPARE
455
BACKGROUND REFERENCE
450
BLEND WITH BACKGROUND
410
FIRST THRESHOLD
YES
420
MESSAGE TO BEHAVIOR MODULE
NO
NO MOVING OBJECTS MESSAGE
415
FRAME HISTORY LOGGING
430
INITIALIZE FRAME HISTORY LOG
435
FRAME EVERY 10 SECONDS
440
COMPARE
445
SECOND THRESHOLD
425

FIG. 5

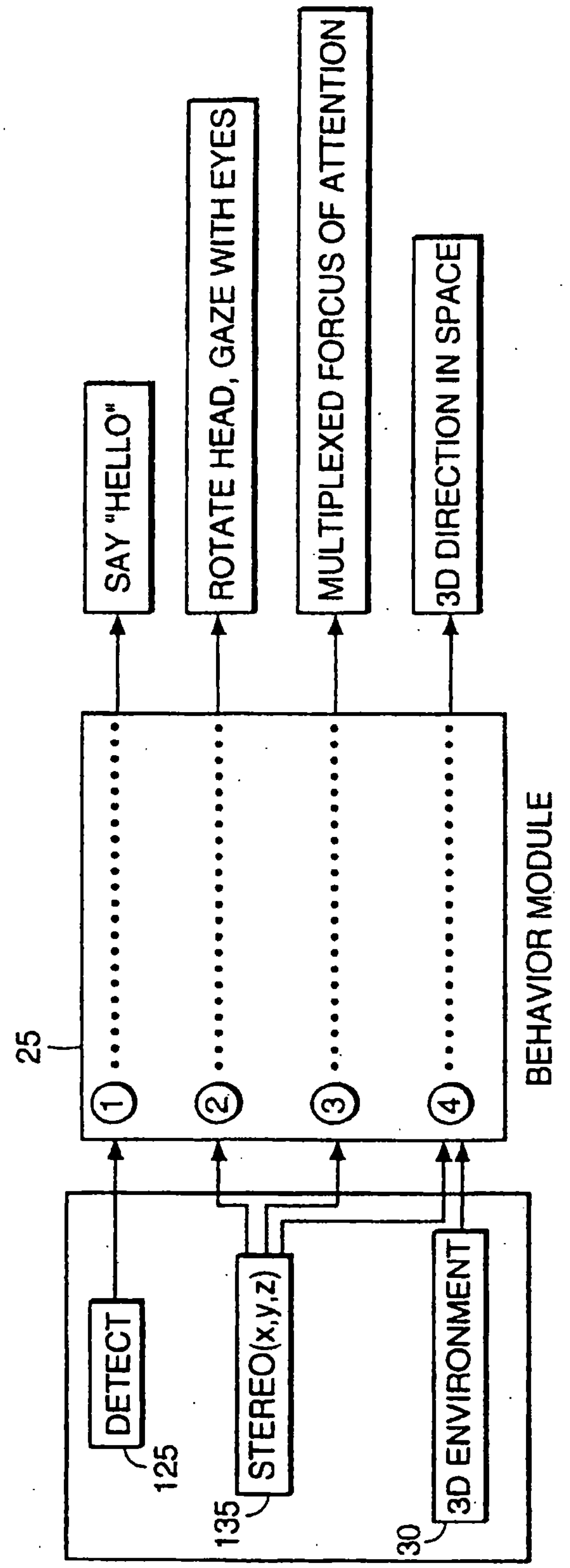

SAY "HELLO"
ROTATE HEAD, GAZE WITH EYES
MULTIPLEXED FORCUS OF ATTENTION
3D DIRECTION IN SPACE
25
1
2
3
4
BEHAVIOR MODULE
DETECT
125
STEREO(x,y,z)
135
3D ENVIRONMENT
30

FIG. 6

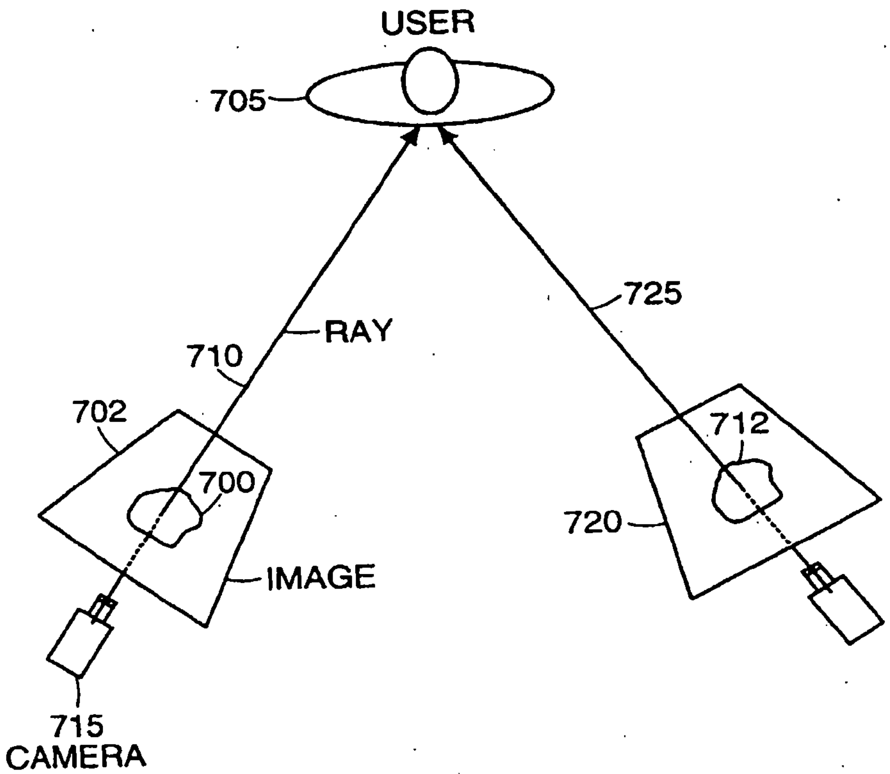


FIG. 7

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 98 10 6380

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
X Y A	EP 0 582 989 A (ISTITUTO TRENTINO DI CULTURA) 16 February 1994 * page 3, line 34 - page 4, line 51 * * page 5, line 15-32 * * figure 1 *	1,7,8,18 2,4,5,14 6,10-12	G06K11/08
X A	US 5 247 433 A (KITAURA WATARU ET AL) 21 September 1993 * column 13, line 36 - column 14, line 30 * * figure 20 *	1,7,8,18 11	
Y A	PATENT ABSTRACTS OF JAPAN vol. 097, no. 008, 29 August 1997 & JP 09 097337 A (FUJI HEAVY IND LTD), 8 April 1997 * abstract *	2,14 1,6-8,12,18	TECHNICAL FIELDS SEARCHED (Int.Cl.6)
Y A	EP 0 694 833 A (AT & T CORP) 31 January 1996 * page 3, line 45-55 * * page 4, line 11-29 * * page 7, line 37-42 * * figures 1-3 *	4,5 9,10,15	G06K G06F
A	"METHOD FOR EXTRACTING FACIAL FEATURES BY USING COLOR INFORMATION" IBM TECHNICAL DISCLOSURE BULLETIN, vol. 38, no. 10, 1 October 1995, pages 163-165, XP000540455 * the whole document *	3,17	

-/--

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	31 July 1998	Baldan, M

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 98 10 6380

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
X,P	EP 0 818 726 A (SIEMENS AG) 14 January 1998 * column 3, line 52 - column 5, line 48 * * column 6, line 5-47 * * figures 1,2,3B-4B *	1,7,8,18	
A		10,11	

TECHNICAL FIELDS SEARCHED (Int.Cl.6)

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	31 July 1998	Baldan, M

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

European Patent Office
Postbus 5818
2280 HV RIJSWIJK
NETHERLANDS
Tel.: +31 70 340 2040
Fax: +31 70 340 3016

Europäisches Patentamt | European Patent Office | Office européen des brevets

DeVile, Jonathan Mark
D Young & Co
120 Holborn
London EC1N 2DY
GRANDE BRETAGNE

EPO Customer Services

Tel.: +31 (0)70 340 45 00

Date
04.05.07

Reference	Application No./Patent No.
P009621EP	00903984.3 - 1224 PCT/JP0000872
Applicant/Proprietor	
Yugen Kaisha GM & M	

COMMUNICATION

The European Patent Office herewith transmits as an enclosure the supplementary European search report under Article 157(2)(a) EPC for the above-mentioned European patent application.

If applicable, copies of the documents cited in the European search report are attached.

☐ Additional set(s) of copies of the documents cited in the European search report is (are) enclosed as well.

Refund of the search fee

If applicable under Article 10 Rules relating to fees, a separate communication from the Receiving Section on the refund of the search fee will be sent later.

Europäisches Patentamt • European Patent Office • Office européen des brevets

European Patent Office

SUPPLEMENTARY EUROPEAN SEARCH REPORT

Application Number
EP 00 90 3984

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
Y	US 5 839 109 A (IWAMIDA HITOSHI [JP]) 17 November 1998 (1998-11-17) * figures 1-6 * * column 1, line 8 - line 14 * * column 1, line 48 - line 56 * * column 2, line 19 - line 59 * * column 3, line 14 - column 4, line 25 *	1,3-25	INV. H04R3/00 H04R25/00
Y	GB 2 256 959 A (WALLACE NIGEL GLYN [GB]; WALLACE NIGEL GLYN [GB]) 23 December 1992 (1992-12-23) * the whole document *	1,3-25	
A	WO 97/29482 A (UNIV CALIFORNIA [US]) 14 August 1997 (1997-08-14) * figure 8 * * page 7, line 9 - page 8, line 4 * * page 17, line 32 - page 20, line 30 * * page 27, line 4 - line 33 * * page 95, line 16 - page 96, line 28 *	1-25	
			TECHNICAL FIELDS SEARCHED (IPC)
A	US 5 326 349 A (BARAFF DAVID R [US]) 5 July 1994 (1994-07-05) * the whole document *	1-25	G10L G09B A61F
A	US 5 283 833 A (CHURCH KENNETH W [US] ET AL) 1 February 1994 (1994-02-01) * the whole document *	7,12-14	G06K G06F H04M H04N
A	EP 0 872 808 A1 (DIGITAL EQUIPMENT CORP [US] HEWLETT PACKARD DEVELOPMENT CO [US]) 21 October 1998 (1998-10-21) * the whole document *	9-11	H04R

-/--

The supplementary search report has been based on the last set of claims valid and available at the start of the search.

3

Place of search	Date of completion of the search	Examiner
The Hague	20 April 2007	Moscu, Viorel

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C04)

European Patent Office

SUPPLEMENTARY EUROPEAN SEARCH REPORT

Application Number
EP 00 90 3984

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
A	US 4 425 481 A (MANSGOLD STEPHAN [SE] ET AL MANSGOLD STEPHAN [SE] ET AL) 10 January 1984 (1984-01-10) * the whole document * * column 1, line 5 - line 27 * * column 2, line 15 - line 22 * * column 2, line 41 - line 50 * -----	1-25	
			TECHNICAL FIELDS SEARCHED (IPC)

The supplementary search report has been based on the last set of claims valid and available at the start of the search.

Place of search	Date of completion of the search	Examiner
The Hague	20 April 2007	Moscu, Viorel

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

3

EPO FORM 1503 03.82 (P04C04)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO.

EP 00 90 3984

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report.
The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

20-04-2007

Patent document cited in search report		Publication date	Patent family member(s)			Publication date
US 5839109	A	17-11-1998	JP	7084592	A	31-03-1995
GB 2256959	A	23-12-1992	NONE			
WO 9729482	A	14-08-1997	EP	0880772	A1	02-12-1998
			JP	2000504849	T	18-04-2000
			US	6542857	B1	01-04-2003
			US	5729694	A	17-03-1998
US 5326349	A	05-07-1994	CA	2139776	A1	20-01-1994
			EP	0651626	A1	10-05-1995
			JP	8501950	T	05-03-1996
			WO	9401059	A1	20-01-1994
US 5283833	A	01-02-1994	NONE			
EP 0872808	A1	21-10-1998	DE	69832119	D1	08-12-2005
			DE	69832119	T2	03-08-2006
			JP	11053083	A	26-02-1999
			US	6256046	B1	03-07-2001
US 4425481	A	10-01-1984	AU	557591	B2	24-12-1986
			AU	8264782	A	21-10-1982
			CA	1176366	A1	16-10-1984
			DE	3268232	D1	13-02-1986
			DK	168582	A	17-10-1982
			EP	0064042	A1	03-11-1982
			JP	6083517	B	19-10-1994
			JP	57185800	A	16-11-1982
			SE	428167	B	06-06-1983
			SE	8102466	A	17-10-1982

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

G0230 SP0137 04.10/29 4回目

整理番号:99P540T01 発送番号:317038 発送日:平成16年 8月30日 1

拒絶理由通知書

FILE	G-610-1
管理	992766
DUE	OCT. 29. 2004
伊藤貴	

特許出願の番号　平成１１年　特許願　第３３８４５８号

起案日　平成１６年　８月２６日

特許庁審査官　石丸　昌平　　９５５９　５Ｃ００

特許出願人代理人　稲葉　良幸（外　５名）　様

適用条文　第２９条第２項、第３６条

<<<<　最　後　>>>>

この出願は、次の理由によって拒絶をすべきものである。これについて意見があれば、この通知書の発送の日から６０日以内に意見書を提出して下さい。

理　由

○平成１６年６月２２日に行った面接に基づいて提出された、平成１６年８月２５日付の補正書案（請求項１３－２５）について検討する。

１．この出願の下記の請求項に係る発明は、その出願前日本国内において頒布された下記の刊行物に記載された発明に基いて、その出願前にその発明の属する技術の分野における通常の知識を有する者が容易に発明をすることができたものであるから、特許法第２９条第２項の規定により特許を受けることができない。

２．この出願は、特許請求の範囲の記載が下記の点で、特許法第３６条第６項に規定する要件を満たしていない。

記　　（引用文献等については引用文献等一覧参照）

［引用文献等一覧］

ア．特開平０３－０３５２９６号公報

１．特開平０５－２８９６０８号公報

２．実願昭５９－１９７８１０号（実開昭６１－１１４４７２号公報）のマイクロフィルム

３．特開平０７－０１３５８２号公報

４．特開昭５１－０５５５０４号公報

５．特開平０８－０７９８９７号公報

６．特開平０２－０９７２００号公報

７．特開平０７－３２７２１３号公報

G-610-1/992766

2/4

8．特開平０４－２４９９９０号公報
9．特開平０８－０６５６４７号公報
１０．特開平１０－１０８１５２号公報
１１．特開平０９－２９２９７１号公報
１２．特開平１０－２２８３６７号公報
１３．特開平０５－０５６４９９号公報
１４．特開昭６０－１４３１００号公報
１５．特公平０７－０５３１６８号公報
１６．特表平１０－５０８４４２号公報
１７．特開昭６２－２２４３４９号公報
１８．特公平０６－０４２７６０号公報

（理由１）

［請求項１３－２５について］：引例１－１８、引例ア

出願人は補正書案において、旧請求項１－１２を削除するとともに、旧請求項１３－２５に係る発明について「認識結果に係る音声情報を要約する手段（旧請求項２５に相当）を備える」旨の記載を加えて、構成を限定した。

ところで、引例５ないし引例６には、音声認識手段を備えた補聴器が記載されている。

また、引例アには、入力文字の解析結果を平易な単語に置き換えて音声出力する技術（テキスト音声合成装置）が記載されている。

ここで、引例５ないし引例６、及び引例アは、音声認識技術という共通の技術分野に属し、また、円滑なコミュニケーションを補助するという一般的な課題を有しているものと認める。

よって、引例５ないし引例６に記載の発明に、引例アに記載の「認識した結果を要約する（平易な単語に置き換える）」技術を組み合わせることは、当業者が容易に想起し得たものと認める（請求項１３、２１－２５）。

また、上記以外の部分については実質的な補正がないため、平成１５年１０月２８日起案の拒絶理由通知で示した以下の文献により、拒絶理由がある。

・引例１５及び１６：人工内耳機構を用いる点（請求項１４）
・引例１５及び１７：人工中耳機構を用いる点（請求項１６）
・引例１５：振動や電気信号を用いる点（請求項１５、１８他）
・引例１８：超音波を用いる点（請求項１７）
・引例１５－１７：補聴器をインプラントする点（請求項１９）
・引例１１及び１２：ワイヤレスで装置間のデータ伝送を行う点（請求項２０）
・引例２他：音声情報を表示する点（請求項２３）

したがって、引例１－１８（特に上記各引例）及び引例アに記載の発明及び周知技術等に基づいて本願の補正書案の請求項１３－２５の如く構成することは、

3/4

当業者が容易になし得たものと認める。

（理由２）

［請求項２２について］

平成１５年１０月２８日起案の拒絶理由通知を参照。 音声認識障害

なお、平成１６年８月２５日付の補正書案について以下の通りコメントする。

・請求項１３について

「要約する変換する」（日本語として不自然）

「少なくとも一つ」（当初明細書に斯かる記載はない）

なお、拒絶の理由が新たに発見された場合には拒絶の理由が通知される。また補正を行う場合は、意見書においてその根拠を明示するとともに、必要なら本願発明のポイントや引用発明と比較した有利な効果を主張されたい。

最後の拒絶理由通知とする理由

１．最初の拒絶理由通知に対する応答時の補正によって通知することが必要になった拒絶の理由のみを通知する拒絶理由通知である。

先行技術文献調査結果の記録

調査した分野（ＩＰＣ第７版）

Ｇ１０Ｌ１５／００－１５／２８、Ｇ０９Ｂ２１／００、

Ｈ０４Ｒ２５／００、Ｈ０４Ｎ７／１４

先行技術文献

特開平０７－１８１８８８号公報、特開平０４－１５６０３３号公報

特開平０７－０８４５９２号公報、特開昭６３－２４９５６０号公報

特開昭６２－２３１９８１号公報、特開平０７－１９１５９９号公報

本通知書の内容に関する問い合わせ、本案件について面接を希望する場合は、特許審査第四部映像機器の石丸（電話:03-3581-1101 内6985）までご連絡下さい。

4/4